2023

Development of Machine Learning based approach to predict fuel consumption and maintenance cost of Heavy-Duty Vehicles using diesel and alternative fuels

Sasanka Katreddi
West Virginia University, gk0037@mix.wvu.edu

Development of Machine Learning based approach to predict fuel consumption and maintenance cost of Heavy-Duty Vehicles using diesel and alternative fuels

Sasanka Katreddi

Dissertation submitted to the
Statler College of Engineering and Mineral Resources
at West Virginia University
in partial fulfillment of the requirements for the degree of

Ph.D. in
Computer Science

Natalia A. Schmid, D.Sc., Chair
Arvind Thiruvengadam, Ph.D., Co-Chair
Gianfranco Doretto, Ph.D.
Xin Li, Ph.D.
Vishnu Padmanaban, Ph.D.

Lane Department of Computer Science and Electrical Engineering

Morgantown, West Virginia
2023

Keywords: Artificial Intelligence, Machine Learning, Neural Networks, Ensemble Models, Mixed Effects Models, Heavy-Duty Vehicles, Fuel Consumption, Maintenance and Repair Cost

Abstract

Development of Machine Learning based approach to predict fuel consumption and maintenance cost of Heavy-Duty Vehicles using diesel and alternative fuels

Sasanka Katreddi

One of the major contributors of human-made greenhouse gases (GHG) namely carbon dioxide (CO_2), methane (CH_4), and nitrous oxide (NO_X) in the transportation sector and heavy-duty vehicles (HDV) contributing to about 27% of the overall fraction. In addition to the rapid increase in global temperature, airborne pollutants from diesel vehicles also present a risk to human health. Even a small improvement that could potentially drive energy savings to the century-old mature diesel technology could yield a significant impact on minimizing greenhouse gas emissions. With the increasing focus on reducing emissions and operating costs, there is a need for efficient and effective methods to predict fuel consumption, maintenance costs, and total cost of ownership for heavy-duty vehicles. Every improvement so achieved in this direction is a direct contributor to driving the reduction in the total cost of ownership for a fleet owner, thereby bringing economic prosperity and reducing oil imports for the economy. Motivated by these crucial goals, the present research considers integrating data-driven techniques using machine learning algorithms on the historical data collected from medium- and heavy-duty vehicles.

The primary motivation for this research is to address the challenges faced by the medium- and heavy-duty transportation industry in reducing emissions and operating costs. The development of a machine learning-based approach can provide a more accurate and reliable prediction of fuel consumption and maintenance costs for medium- and heavy-duty vehicles. This, in turn, can help fleet owners and operators to make informed decisions related to fuel type, route planning, and vehicle maintenance, leading to reduced emissions and lower operating costs.

Artificial Intelligence (AI) in the automotive industry has witnessed massive growth in the last few years. Heavy-duty transportation research and commercial fleets are adopting machine learning (ML) techniques for applications such as autonomous driving, fuel economy/emissions, predictive maintenance, etc. However, to perform well, modern AI methods require a large amount of high-quality, diverse, and well-balanced data, something which is still not widely available in the automotive industry, especially in the division of medium- and heavy-duty trucks. The research methodology involves the collection of data at the West Virginia University (WVU) Center for Alternative Fuels, Engines, and Emissions (CAFEE) lab in collaboration with fleet management companies operating medium- and heavy-duty vehicles on diesel and alternative fuels, including compressed natural gas, liquefied propane gas, hydrogen fuel cells, and electric vehicles. The data collected is used to develop machine learning models that can accurately predict fuel consumption and maintenance costs based on various parameters such as vehicle weight, speed, route, fuel type, and engine type.

The expected outcomes of this research include 1) the development of a neural network model that can accurately predict the fuel consumed by a vehicle per trip given the parameters such as vehicle speed, engine speed, and engine load, and 2) the development of machine learning models for estimating the average cost-per-mile based on the historical maintenance data of goods movement trucks, delivery trucks, school buses, transit buses, refuse trucks, and vocational trucks using fuels such as diesel, natural gas, and propane. Due to large variations in maintenance data for vehicles performing various activities and using different fuel types, the regular machine learning or ensemble

models do not generalize well. Hence, a mixed-effect random forest (MERF) is developed to capture the fixed and random effects that occur due to varying duty-cycle of vocational heavy-duty trucks that perform different tasks. The developed model helps in predicting the average maintenance cost given the vocation, fuel type, and region of operation, making it easy for fleet companies to make procurement decisions based on their requirement and total cost of ownership. Both the models can provide insights into the impact of various parameters and route planning on the total cost of ownership affected by the fuel cost and the maintenance and repairs cost.

In conclusion, the development of a machine learning-based approach can provide a reliable and efficient solution to predict fuel consumption and maintenance costs impacting the total cost of ownership for heavy-duty vehicles. This, in turn, can help the transportation industry reduce emissions and operating costs, contributing to a more sustainable and efficient transportation system. These models can be optimized with more training data and deployed in a real-time environment such as cloud service or an onboard vehicle system as per the requirement of companies.

Acknowledgements

My Ph.D. Journey has been full of development with acquiring knowledge, facing challenges and moving forward which would not have been possible without support of everyone. I would like to express my deepest gratitude to my research advisor, Dr. Arvind Thiruvegadam, who guided me throughout my Ph.D. journey. This dissertation would not have been possible without his invaluable assistance, intellectual support, and encouragement. Besides my advisor, I would like to extend my sincere thanks to my Ph.D. committee chair, Dr. Natalia Schmid for her insightful comments, suggestions, and assistance at every stage of my research. Their expertise and encouragement have been instrumental in shaping my research and ensuring success.

I am thankful to my committee members Dr. Gianfranco Doretto, Dr. Xin Li, and Dr. Vishnu Padmanaban for their encouragement, technical discussions, insightful feedback, and constructive criticism. Their valuable insights have greatly enriched the quality of my research.

Immense thanks to my lab mates Dr. Satish Guda, Batishahi Selimi, Chakradhar Reddy Vardhireddy, Hemanth Kumar, Isacc Boyce, Sai Swarup Noone, and Ramana Kumari Inturi for their help and knowledge sharing. A special thanks to my childhood friend Dr. Sujan Kasani for his continuous encouragement and valuable inputs.

Finally, I would like to thank my father Rama Kumar, mother Sailaja, sister Priyanka and brother-in-law Chandu Gudise for their unconditional love, support, and understanding. Their constant encouragement and belief in me have been a driving force behind my academic achievements.

I am truly grateful to all those who have helped me in any way to complete my dissertation. Without their support and encouragement, this work would not have been possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The transportation sector is one of the major contributors to greenhouse gas emissions contributing about 27% (shown in Figure 1.1) of overall emissions in the United States. Among the transportation sector emissions, medium- and heavy-duty vehicles produce 26% as per 2020 reports even though they only contribute 4% of vehicles on road [1]. The increasing greenhouse gas emissions result in global warming adversely impacting human health, the environment, and the economy.

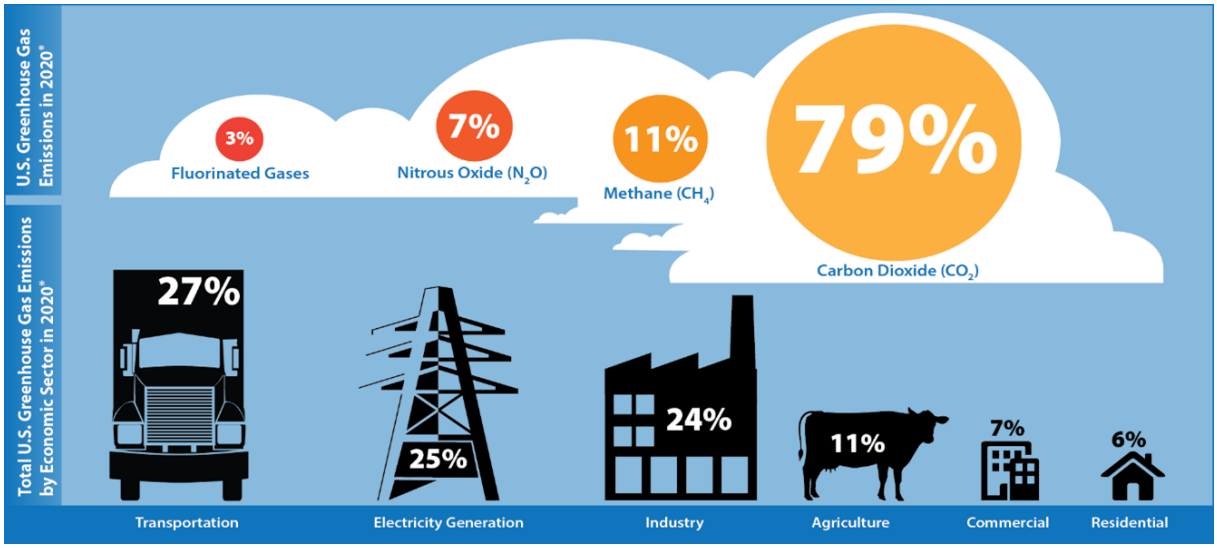

Figure 1.1: Sources of United States Greenhouse Gas Emissions in 2020 [1]

However, despite these effects road transportation of goods is increasing every day resulting in

high energy consumption. The transportation sector alone consumes 28% of the energy consumed in the United States (as shown in Figure 1.2) [2]. Improving the fuel efficiency of vehicles reduces greenhouse gas emissions and increases energy security due to decreased fuel consumption. Considering the adverse effects on climate and the high energy usage, the U.S. environmental protection agency (EPA) and national highway traffic safety administration (NHTSA) proposed fuel efficiency and greenhouse gas emission standards for commercial vehicles. Under the phase 1 regulation, vocational vehicles including goods movement, delivery trucks, school buses, and refuse trucks are required to reduce fuel consumption and greenhouse gas (GHG) emissions by 10% by 2018. Hence, US EPA introduced corporate average fuel economy (CAFÉ) standards for manufacturers to be compliant with standards to reduce fuel consumption and minimize exhaust pollution.

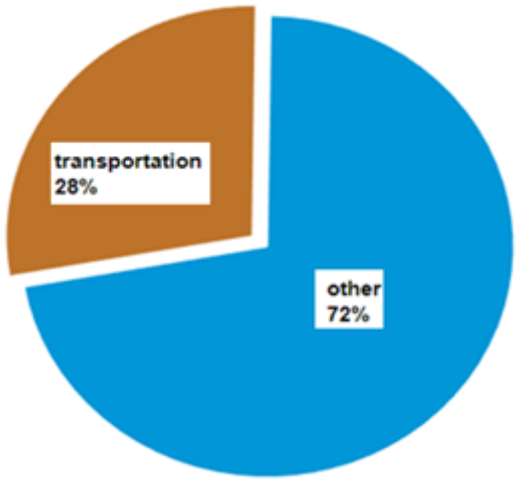

Figure 1.2: Total U.S. energy used for Transportation in 2021 [2]

Not only the higher emissions but the higher energy consumption also increases the total cost of ownership (TCO) for fleet management companies. About 24% of cost-per-mile for a commercial truck is spent on fuel and 7% on maintenance and repairs. Thus studying the factors that influence the fuel consumption in heavy-duty vehicles and seeking low-carbon alternative fuel vehicles that use natural gas, propane, hydrogen fuel cells, or electric vehicles could help in taming the emissions in the long term. Reducing fuel consumption and emissions from existing internal combustion engine vehicles is the short-term goal [10] for a sustainable future. This in turn reduces the cost-per-mile for a commercial truck.

1.2 Objectives

The objective of this work is to study fuel consumption and maintenance and repairs which are the two factors that influence the total cost of ownership in heavy-duty vehicles using machine learning. Machine learning coined under artificial intelligence uses algorithms and neural network models to progressively improve performance. These models apply historical data to understand the patterns in heavy-duty vehicles (HDVs) to be able to predict new data for which the classification or the output is unknown without using on-road testing or heavy equipment. The main concepts covered in this dissertation are developing/using machine learning algorithms to model real-world on-road heavy-duty vehicle data. The objectives are listed below:

1. To develop a data-based machine learning model for estimating the fuel consumption of a trip in conventional diesel heavy-duty trucks using very few engine parameters that can be easily obtained.

2. To investigate the patterns of maintenance cost in medium- and heavy-duty vehicles using machine learning algorithms and develop a model for estimating the average cost-per-mile for diesel, natural gas, and propane heavy-duty vehicles.

3. To develop a single generalized model for estimating the maintenance cost of diesel and alternative fuel (natural gas and propane) medium- and heavy-duty trucks performing various activities such as goods movement, delivery trucks, school buses, transit buses, refuse trucks, and vocational trucks.

1.3 Significance

Machine learning uses historical data and statistical analysis to capture the patterns in data automating the decision-making process. But the collection of real-world road data from vehicles meeting all standards and requirements is challenging. Due to the lack of availability of data most of the existing studies are limited because they are confined to using synthetic data or data collected in a lab under certain circumstances. The work presented in this dissertation is significant in

1. developing a neural network model by identifying and using very few vehicle parameters that affect fuel consumption most and can be easily obtained from vehicle trip data;

2. collecting real-time medium- and heavy-duty vehicle maintenance data in collaboration with fleet management companies;

3. modeling maintenance data for predicting maintenance cost of diesel, natural gas, and propane-fueled heavy- and medium-duty vehicles that include goods movement, delivery trucks, school buses, transit buses, refuse trucks, and vocational trucks that enable companies to make procurement decisions.

The developed machine learning models are optimized using real-time data collected regularly improving the prediction accuracy. The models can be deployed in a real-time environment such as a cloud service or an onboard vehicle system. This involves integrating the model with the rest of the system into a framework that gives real-time insights.

1.4 Thesis Organization

The document includes seven chapters. The first chapter contains the introduction, objective, and significance of the thesis. The second chapter details the background of Artificial Intelligence, Machine Learning, Deep Learning, and applications. Chapters 3, 4, 5, and 6 are expanded versions of published or submitted manuscripts. Chapter 3 discusses the application of artificial intelligence in heavy-duty trucks describing existing studies and identifying the research gaps. Chapter 4 focuses on identifying the vehicle features that impact fuel consumption most and the estimation of fuel consumption in heavy-duty vehicles using real data collected at the West Virginia University Center for Alternative Fuels Engines and Emissions (WVU CAFEE). Chapter 5 presents machine learning algorithms for estimating maintenance and repair costs in delivery trucks. Chapter 6 describes the estimation of maintenance cost in vocational trucks with alternative fuels using an ensemble machine learning model and showcases the importance of this study. Chapter 7 summarizes the contributions of this work and gives a brief outline of future work.

Chapter 2

Background

The Fourth Industrial Revolution or Industry 4.0 [11] has emerged with the goal of intelligent, inter connective, and automation technologies. Artificial Intelligence (AI) is one of the technologies emerging as part of Industry 4.0 technologies including smart manufacturing, big data analytic, cloud, Internet of Things (IoT), block chain, and simulation. AI is creating intelligent and smart machines that can exhibit natural intelligence like humans. AI makes it possible for machines to learn like humans from experience and change behavior based on inputs. A typical AI analyzes its environment and takes actions that maximize its chance of success. In recent years elevated interest in AI and the latest advances in computer technologies led to the rapid growth of Machine Learning (ML) providing us with algorithmic means for the intelligent processing and analysis of large sets of data. Machine Learning (ML) and Deep Learning (DL) as a subset of AI and the rise of AI is shown in Figure 2.1.

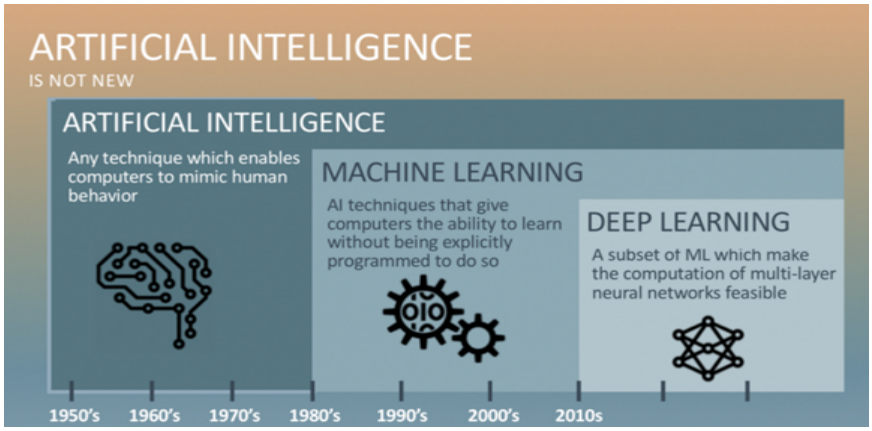

Figure 2.1: Machine Learning and Deep Learning as part of Artificial Intelligence [3]

2.1 Machine Learning

In 1959, Arthur Samuel defined Machine Learning as "The field of study that gives computers the ability to learn without being explicitly programmed" [12]. Machine Learning evolved as a subset of Artificial Intelligence where machine learning feeds on computer data and uses a statistical technique to help it learn how to progressively get better without having been programmed. By learning from data, the algorithms can build predictive models by identifying patterns in data. The goal of Machine Learning is to achieve good predictions enough to be useful but not to achieve perfect predictions. However, the performance of the machine learning algorithm depends on the characteristics and complexity of the data [5]. The typical workflow for machine learning is shown in Figure 2.2.

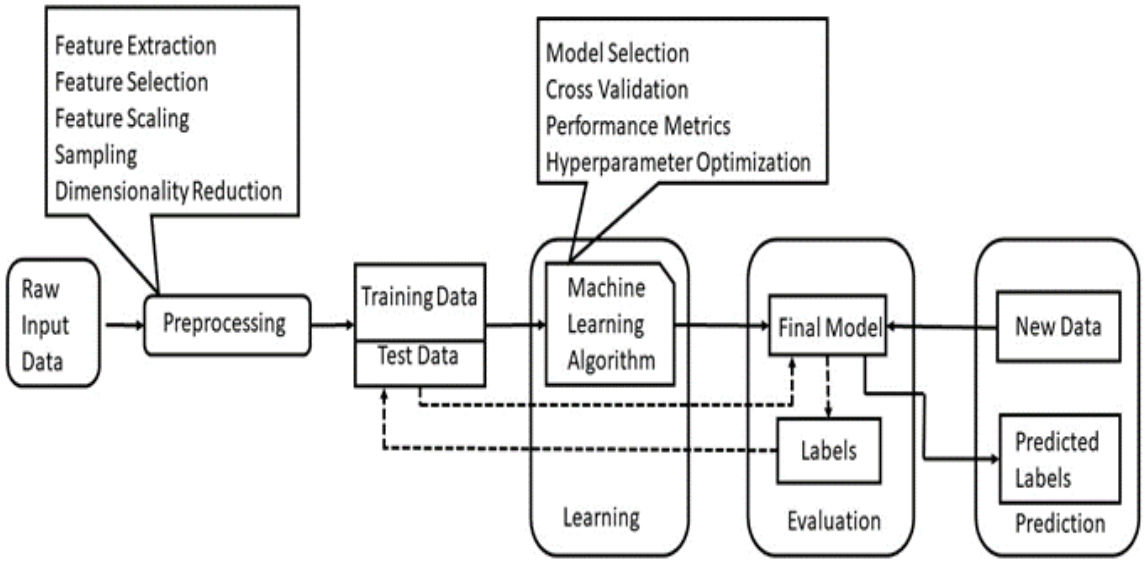

Figure 2.2: Machine Learning Workflow [4]

The subsets of machine learning such as supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning are based on the type of data. The algorithms can be fed with labeled or unlabeled data based on which type of learning is determined. Thus, the key to machine learning is data, the amount of training is based on the amount of data, and the amount of training the system should undergo is determined by the type of algorithm. Machine-Learning implementations are classified into four major categories as shown in Figure 2.3 which are discussed in Section 2.1.1 depending on the nature of the signal and response.

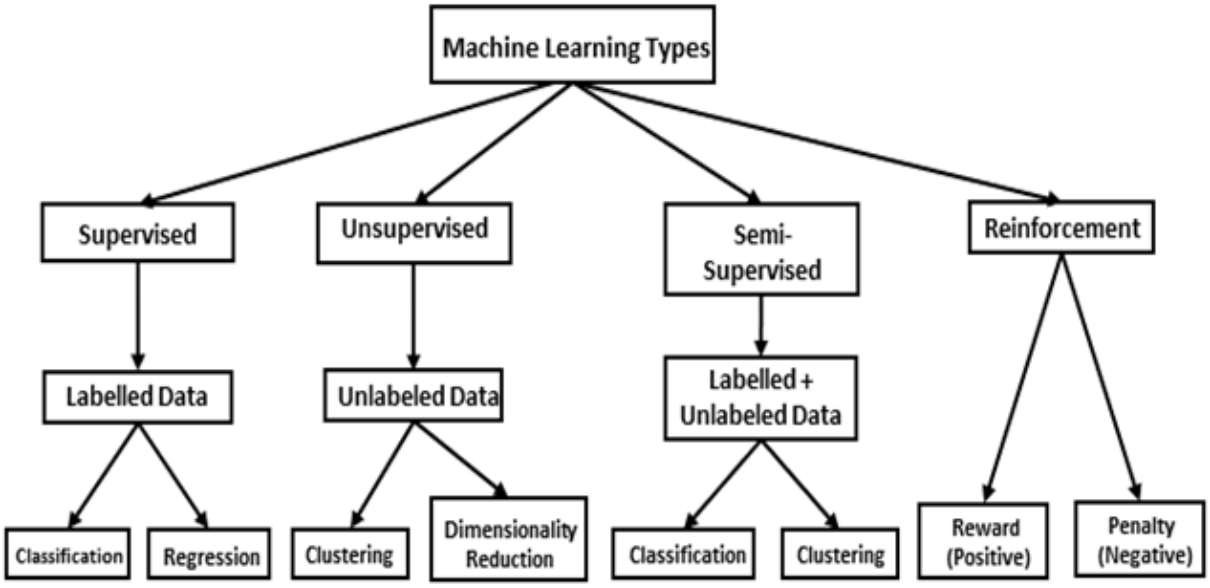

Figure 2.3: Machine Learning Types [5]

2.1.1 Types of Machine Learning Techniques

1. Supervised Learning: In supervised learning, the machine learning algorithms learn from mapping input data and its associated ground truth by developing a model. This model later predicts the response for new data samples. This is more like training humans by showing examples. Hence this method of learning uses labeled data. The most common supervised tasks are classification and regression.

2. Unsupervised Learning: In unsupervised learning, the algorithm learns from input data but without any ground truth information. The algorithm learns from the patterns within data leading to new features that represent the class for each sample. This technique is used for identifying trends in data, data exploration, and extracting generative features. The most common unsupervised learning tasks are clustering, dimensionality reduction, feature extraction, anomaly detection, etc.

3. Semi-Supervised Learning: As the term, Semi indicates this is a hybridization of supervised and unsupervised techniques. The learning process for this type of algorithm is not closely supervised and does not allow it to do things on its reducing burden. The dataset is designed in such a way that it has more unlabeled data and a small, labeled dataset. The availability of

labeled data in most scenarios could be limited, where semi-supervised learning is beneficial. Semi-supervised learning is used in fraud detection, text classification, labeling data, and machine translation.

4. Reinforcement Learning: Reinforcement learning is mostly a trial-and-error-based approach as algorithms do not take labels for learning but take positive and negative feedback which reinforces the algorithm. It is an environment-driven approach based on reward and penalty using insights from environmental activists to take action to increase the reward and minimize the risk. This technique is used to train models that can help increase automation or optimize performance such as autonomous driving, robotics, manufacturing, and supply chain.

Machine Learning algorithms are designed to learn an unknown function (target function, f) from possibly available training data. This target function maps features (input, x) to an output variable (y) resulting in y=f(x)+e relation, where e is the error. Once f(x) is designed, it is applied to testing data to evaluate the performance of the designed function.

2.1.2 Benefits of Machine Learning

1. Easily adaptable to data changes.
2. The decision-making process is given priority resulting in fast decision making
3. Data patterns can be analyzed easily, and action can be taken based on them.

Since the time of its introduction, ML grew in a large scope populated with methods, algorithms, and techniques. However, there is always room for more. Due to the latest breakthrough in computer technology yielding parallel data processing, ML techniques increased their demand for memory leading to the introduction of Deep Learning.

2.2 Deep Learning

Deep Learning is a Machine Learning technique that runs inputs through several layers connected by neurons like structures making it capable of learning from data. Neural Networks can learn from raw data with minimal preprocessing and the accuracy of prediction increases due to

flexible architectures. The number of hidden layers determines how deep the machine can learn by making connections. The best results can be achieved by minimizing loss through adjusting weights and biases. The history of Deep Learning dates to 1943 when Walter Pitts and Warren McCulloch created a computer model based on the human brain. 1985-the 90s is the second AI winter, which affected research for neural networks and deep learning. In 1999, when a computer with GPUs was developed the neural networks started to compete with Support Vector Machines (SVM). With the increased computational power, neural networks have gained popularity recently [13]. Deep Learning mainly handles complex mappings from input to output and requires large datasets and high computational power.

2.2.1 Types of Neural Networks

1. Perceptron Model: The simplest and oldest Neural Network with a single neuron [**?**].

2. Multi-Layer Perceptron or MLP: A Neural Network like perceptron but with multiple hidden layers adding complexity. Most of the neurons are present in the hidden layers making it the key for data manipulation to achieve the required output. The output layer also contains neurons that determine the object. They work on the principle of feed-forward, hence they are also called feed-forward networks. These networks are mostly used in Computer Vision and Speech Recognition applications [14].

3. Convolutional Neural Network (CNN): An MLP with convolutional blocks for analyzing images and videos. A convolutional filter of size NxM is moved across the images to pick hidden features of signals or objects that may be of interest to us. The most important features are identified using pooling layers. Convolutional Networks are widely used in Image Classification, Computer Vision, and Pattern Recognition [15, 16].

4. Recurrent Neural Network (RNN): RNN analyzes other states of data like the data in past and predicts the future. The state matrices of RNN stores the last output. RNN uses the concepts of both MLP and previous outputs. These networks are used in Stock Market Predictions, Time Series Data Predictions [17].

5. Autoencoders (AE): Rather than taking decisions these network works to maintain quality

even though the data are compressed. The hidden layer in autoencoders is called a bottleneck which is a smaller layer than the input and output making the network compress data but maintain quality.

There are many more neural networks like Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Sparse AE (SAE), Variational AE (VAE), Denoising AE (DAE), etc as shown in Figure 2.4 (adapted from [16]). They are tuned to work with various tasks and applications.

Backpropagation Algorithm: Backpropagation is a training algorithm that propagates the error back for training weights in a Multi-Layer Perceptron. The backpropagation calculates error by comparing the output with ground truth values and adjusting the weights. The back-error propagation takes place layer by layer by adjusting the weights.

2.2.2 Benefits of Deep Learning

1. No need to extract features from raw data i.e. DL extracts the best features for us. Deep Learning scans for features that correlate and combine them to enable faster learning.
2. Maximum utilization of unstructured data.
3. High-quality results can be achieved.
4. Eliminates the need for data labeling and unnecessary costs.

2.3 Applications of AI

AI is proving its ability in solving daily complex problems in many industries efficiently. With recent advancements in big data and the computational power of computers, AI is advancing its ability in reasoning, planning, knowledge, communication, etc. From voice assistants like Siri to Self-driving vehicles, AI has revolutionized technology with its advancements. The voice assistants like Siri and Alexa use machine learning algorithms to become smarter and understand natural speech. Autonomous vehicles are becoming smarter and smarter day by day with specialized features like auto braking systems, self-parking, etc. [18]. Searching for anything online has become much easier as companies like Amazon, Netflix, Pandora, etc. are using AI to provide the best suggestions for their customers based on previous data [19]. AI is being used in the healthcare sector

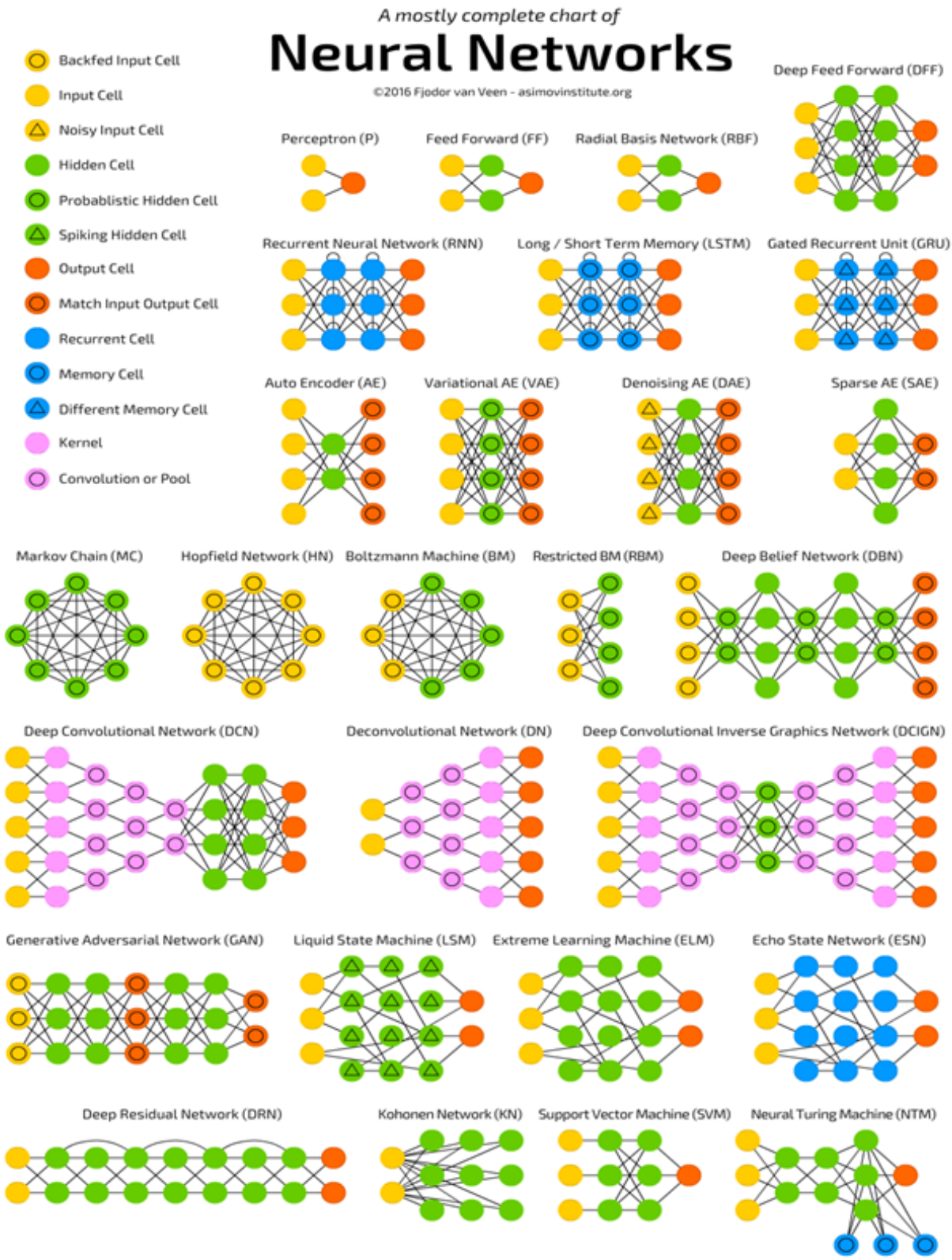

Figure 2.4: Types of Neural Networks [6]

for fast diagnosis than doctors. It is being possible to predict the chances of diseases ahead [20]. Natural Language Processing has made it possible for humans to talk with computers, and smartphones [21]. Social Media which has become a part of daily life for everyone is curated by AI from notifications to feeds [22, 23]. Many more applications have powerful AI applications shown in Figure 2.5.

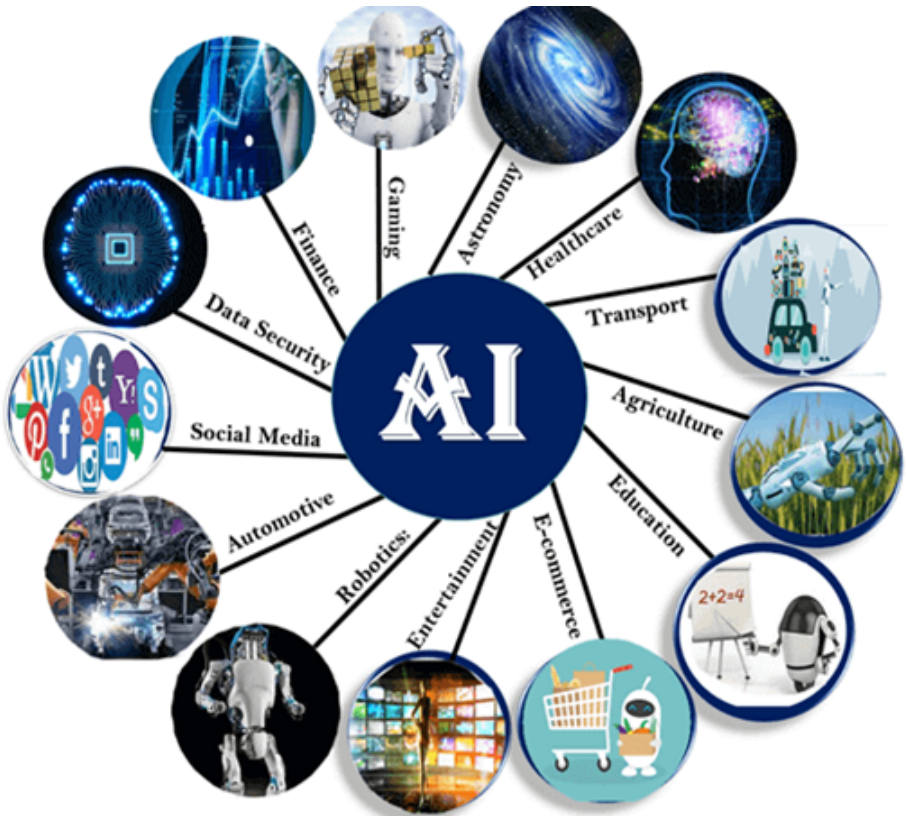

Figure 2.5: Applications of AI in some sectors [7]

In this proposal, several ML techniques are applied to one-dimensional tabular data collected from Heavy-Duty Vehicles (HDVs). The success of deep learning on tabular data is not yet clear even though it has seen tremendous success on images, videos, and text datasets [24]. With the increasing number of vehicles on road, there is also an increase in natural resource consumption, increased emissions, and reduced safety due to more accidents. AI has already proven its success in automobiles through self-driving cars and many other applications. However, there is a gap in extending this application to heavy-duty vehicles. With the increased e-commerce, there have been more trucks on highways these days transporting goods. Heavy-duty vehicles using fuel are

predominantly in use which in turn increases fuel consumption and emissions. Many internal, external, and environmental factors affect the fuel consumption, emissions, and maintenance of vehicles. It is difficult for the human brain to understand the trends in huge volumes of data collected from vehicles. Hence machine learning algorithms can make the analysis easier. Deep learning, a class of machine learning algorithms has gained success on data containing images, videos, text, audio, etc. However, on tabular data, neural network performance is still unclear and machine learning ensemble models remain the promising tool as they don't need differentiable functions, and creating inductive biases required for neural networks on tabular data made of heterogeneous features and the small sample size is difficult. Neural networks on tabular data require a lot of data pre-processing and exploratory data analysis as data containing irrelevant features makes the neural network performance worse especially with irregular patterns in the target. Most importantly, Machine learning algorithms are easy to train and have low tuning costs. In this work, various machine learning algorithms for predicting continuous variables are investigated.

Chapter 3

A Review on Application of Artificial Intelligence in Heavy Duty Trucks

Artificial Intelligence (AI) has seen success in industries such as banking, healthcare, manufacturing, agriculture, transportation, automobile/automotive, and many more. Automotive Indistry ecosystem is being highly impacted and changed by the potential of AI. The customer's preference for advanced and new features, driver's assistance, self-driving, etc. are propelled the usage of AI in the automotive industry. AI is being used in every phase of automotive, from autonomous driving to manufacturing to supply chain to production to driver safety. As per the United States Environmental Protection Agency (U.S. EPA) [25], about 32% of on-road NO_X (nitrous oxide) emissions and 23% of GHG emissions are produced by heavy-duty trucks impacting the climate and people's health. As per National Highway Traffic Safety Administration (NHTSA) reports in 2020, 76% of fatal crashes involved large trucks. Heavy-duty vehicles constitute about 18% of energy use and 17% of petroleum use in the United States [26]. As per reports [27], predictive maintenance can save about 8-12%. A small change in transportation emissions and fuel economy, lowering maintenance time and safety of drivers can have outsized effects and significant global impact on climate and natural resources. Due to advanced technologies, there is a wealth of data collected from vehicles these days. Analyzing this data using emerging technologies like Artificial Intelligence helps draw key insights into heavy-duty transportation without on-road testing of vehicles.

Hence an effort has been made to bring together the existing studies on the application of AI in heavy-duty trucks and identify gaps in current research in this paper. Various studies from the

literature on fuel consumption/efficiency, emissions, self-driving, truck platooning, and predictive maintenance have been presented. The studies presented use various AI techniques that help in identifying the pattern and making predictions/decisions. However, the feasibility and versatility of using data-driven methods and artificial neural networks are unclear in some applications such as turbo machines and electrical submersible pumps [28]. Similarly, due to the challenges involved in implementing Artificial Intelligence in the trucking industry and the lack of surveys on studies related to AI in heavy-duty trucks, this paper focuses especially on the applications of AI in heavy-duty trucks. Truck Platooning which is considered partially autonomous is the next big thing in truck transportation that can be achieved through AI techniques. The increased traffic congestion and accidents, stringent regulations for emissions and fuel consumption, and lack of truck drivers and safety have accelerated the application of AI in the trucking industry. Fleet management companies are already adopting AI technologies such as machine learning, deep learning, computer vision, and natural language processing discussed in the following sections to observe the performance of trucks which can help in cost management, reducing downtime, analyzing truck performance, etc. AI helps fleet owners to make a prediction based on the patterns in the previous data, can enable vehicle 2 vehicle (V2V) communication, identify drivers' behavior, and select the route with low fuel consumption.

3.1 Fuel Consumption/Economy

Fuel consumption is one of the important aspects of vehicles, especially Fleet/Heavy-duty vehicles. The Corporate Average Fuel Economy (CAFÉ) standards by National Highway Traffic Safety Administration regulate the fuel economy standards for vehicles. Fuel economy contributes to the key factor in the overall operational cost of vehicles, especially heavy-duty trucks. Increasing fuel efficiency and reducing fuel consumption can save a lot for transportation companies. Several studies have been made for modeling fuel consumption/fuel efficiency using statistical and other approaches. Predicting fuel efficiency can help in managing the fleet and for diagnostic purposes in case of high fuel consumption. The physics-based and statistical approaches to modeling are time-consuming and less accurate compared to machine learning methods. Several studies have been performed to predict fuel consumption in vehicles using machine learning and deep learning

techniques [29–35]. Perrotta et. al. [36] applied three machine learning techniques namely Support Vector Machine (SVM), Random Forest (RF), and Artificial Neural Network (ANN) for the estimation of fuel consumption in trucks based on telematics and Highways Agency Pavement Management System (HAPMS) data. The parameters used in their modeling are gross vehicle weight, road gradient, vehicle speed, average acceleration, % start torque, % end torque, the engine revs at the start of the record, used gear, cruise control, the radius of curvature of the road, road roughness as Longitudinal Profile Variance (LPV) at 3, 10, and 30 m wavelengths, road surface macrotexture and achieved root mean square error of 5.12, 4.64, 4.88 (liters/100km i.e., liters of fuel consumption per 100 km) for SVM, RF, and ANN respectively. The comparison of results indicates the best performance achieved with random forest having RMSE of 4.64/100km and R^2 of 0.87 but SVM and ANN had better accuracy at prediction. The work is limited to features related to engine, vehicle, and road whereas considering other parameters such as climate, and drivers' behavior can improve the results. Katreddi et. al. [37] predicted the fuel consumption based on the input parameters of engine load (%), engine speed (rpm), and vehicle speed (km/h) of heavy-duty trucks using a feed-forward neural network with backpropagation. The model predicts the average fuel consumption by the truck given the input parameters. The predicted fuel consumption with distance is compared with other machine learning techniques linear regression and random forest. It was shown that the MLP achieved the best performance with an RMSE of 0.0025L (fuel consumed in liters) compared to machine learning techniques such as linear regression and random forest. The data used in this study were collected at WVU CAFEE using a portable emissions measurement system (PEMS) device. The study is limited to understanding the effect of the very few features that can be obtained easily unlike other studies where a collection of sensor data that might contain noise is not required. The shortcoming of this study is not considering external factors such as climate, and GPS information. The work has been confined to a single truck which could impact the fuel consumption significantly in some cases but considering the entire trip covered various stages of engine operations. The authors should have considered other parameters and compared the results of the neural network with a different number of hidden layers. A similar study to predict the fuel consumed in mining dump trucks based on payload, loading time, idled while loaded, loaded travel time, empty travel time, and idled while empty using ANN was performed [38]. The study involved analyzing data from 5001 cycles of haulage operations using a feed-forward neural network (6-9-9-1) with

backpropagation. The results revealed that the idle time of dumping trucks significantly impacts fuel consumption. The gap in addressing unnecessary fuel consumption and reducing emissions during idle speed has been addressed. The consideration of idle energy consumption and emissions is important in vocational trucks such as school buses and dumping trucks that make frequent stops. Another study involving mining trucks was done by Soofastaei et. al. [39]. The haulage vehicles are designed to perform well with heavy loads and with greater road grade and resistance. An ANN was used to find the correlation between fuel consumption and the input parameters: truckload, truck speed, and total haul road resistance. A Genetic Algorithm (GA) is then applied to optimize the fuel consumption based on the input parameters and fitness function created by ANN. The study has used a large dataset that generalizes the model well that could give a good prediction on unseen data. Identifying the range of values for the Gross Vehicle Weight and Truck Speed can help in managing the fuel efficiently. Bodell et. al. [40] compared the performance of machine learning algorithms linear regression (LR), K-Nearest Neighbor (KNN), ANN (MLP) with Adam, and ANN (MLP) with SGD (Stochastic Gradient Descent) using simulated and operational data considering road slopes and driver profiles. For the simulated data scenario, ANN with Adam performed better than other methods with a mean square error of 0.026 liters/100km whereas for operational data both the ANN algorithms (MLP with SGD and MLP with Adam) performed with a mean square error of 2.939 liters/100km. This work has been limited to machine learning models and due to challenges in computational power has not been extended to deep learning methods. The fuel consumption in heavy-duty trucks with a combustion engine is affected by the operating points. Hence, the fuel consumption modeling using engine parameters: engine speed, torque, and fuel consumption at different operating points were performed by Wysocki et. al. [41]. Their work evaluated the performance of polynomial regression, K-Nearest Neighbor (KNN), and Artificial Neural Network (ANN) on the exploitation data collected and observed that ANN trained on the 8 input variables (engine speed and torque at initial, 500, 1000, and 1500th millisecond) performed best with less amount of training data. The sensitivity of training data size has been evaluated for various models which is helpful as the machine learning models are dependent on the amount and variation in data but are limited to combustion engines. A neural network model based on distance windows rather than for predicting the average fuel consumption in heavy vehicles was presented by Schoen et. al. [42]. The model used is a Feed-Forward Neural Network trained using input

parameters: the number of stops, time stopped, average moving speed, characteristic acceleration, aerodynamic speed squared, change in kinetic energy, and change in potential energy. The model was evaluated with different time windows of 1km, 2km, and 5km, and the 1km window has the highest accuracy with an accuracy of 0.91 and RMSE (liters/100km) of 0.0132 for fuel consumption prediction. The window size is dependent on the data and application-specific which is considered a limitation of work. In [43], an Explainable Boosting Machine was used to measure the impact of actionable factors on fuel consumption with the data collected from different vehicles (cars to trucks). An algorithm to generate explanations related to the relationship between fuel consumption to fuel factors in the trained models was proposed. The data used in this work is independent of the fuel type and driving behavior which might be considered a limitation and the analysis of hybrid vehicle fuel consumption is considered a future scope. All the previous studies help manufacturers and fleet companies identify various factors affecting fuel consumption/fuel efficiency for diagnostic purposes and cost management.

3.2 Emission Estimation

As emissions regulations for transportation are being imposed, automotive industries especially heavy-duty trucks are focusing on reducing emissions. Emission estimation helps in developing emission inventories and set standards for environmental protection. Due to challenges in physics-based models, companies are using data-driven approaches for estimating emissions and taking necessary actions to reduce the emissions. Previous studies for analysis and estimation of emissions like carbon monoxide (CO), carbon dioxide (CO_2) [44] nitrogen oxides (NO_X) [45], hydrocarbon (HC), and particulate matter (PM) [46, 47] from vehicles using machine learning have proved the ability of AI in emissions data study [48–55]. These studies have motivated the use of Artificial Intelligence techniques in analyzing and estimating emissions in heavy-duty trucks [56].

Pillai et. al. [57] modeled and predicted engine-out nitrogen oxide (NO_X) and tailpipe nitrogen oxide (NO_X) in heavy-duty vehicles using deep neural networks (DNN). Four DNNs to perform supervised regression tasks for estimating Engine-out NO_X and Tailpipe NO_X were developed using data collected from engine dynamometer and chassis dynamometer testing as input. It has been determined that high-accuracy models can be developed with minimal significant engine and

after-treatment input parameters such as SCR inlet and outlet temperature, engine-out NO_X, and exhaust mass flow rate. It has been observed that engine-out NO_X has good prediction accuracy with $R^2 = 0.99$ whereas the tailpipe NO_X has a prediction accuracy of $R^2 = 0.92$. The results of actual and predicted Engine Out NO_X and Tailpipe NO_X using randomly selected input data. Extending this work on the on-road testing data might be more accurate as it considers the telemetric data. An ANN with Levenberg-Marquardt (LM) training algorithm is used by Mohammadhassani et. al. [58] for NO_X emission prediction of heavy-duty diesel engines. The model takes engine speed, air intake temperature, and mass fuel as inputs and achieved an R^2 of 0.89 for the test data. Considering just the engine operating parameters of specific engine types in heavy-duty vehicles limits the study as the emissions are dependent on other factors such as vocation type, fuel type, age, road conditions, etc. Super Learner [59] model based on Random Forest, XGBoost, Light GBM, and CatBoost was proposed by Wei et.al. [9] for predicting CO_2 and NO_X emissions. One-level (Level 1 Super Learner Regressions) model prediction was employed for CO_2 and two-level (Level 1 Super Learner Regression, Level 2 Super Learner Classification) prediction for NO_X prediction. The super learner achieved an R^2 value of 0.94 and 0.84 for CO_2 and NO_X emissions respectively with the comparison of actual and predicted emissions with other methods. This study focused on onboard test data and was able to predict well for significantly different emissions. A single model that predicts CO_2 and NO_X could be considered. Prediction of emissions from diesel engines which are mainly used in heavy-duty vehicles was studied by Yu et. al. [60] using CEEMDAN-LSTM. The CEEMDAN algorithm is used to extract subseries of NO_X emission data at different frequencies. An LSTM neural network is then trained on the subseries data. The performance of CEEMDAN-LSTM was compared with other machine learning models random forest, support vector regression, XGBoost, LSTM, CEEMDAN-RF, CEEMDA-SVR, and CEEMDAN-XGBoost. Including CEEMDAN algorithm data decomposition on data has reduced the sudden changes in data and improved the accuracy of the LSTM neural network. The CEEMDAN-LSTM approach has a better performance compared to others mentioned in the paper with RMSE of 46.11/ppm, and R^2 of 0.98. However, this work did not consider the effect of GIS parameters in emission estimation. The sensor data collected every second contains a lot of noise and is smoothed before feeding into the LSTM neural network helps in stabilizing the model performance which has not been done in most of the studies. Prediction of emissions from HDVs

based on various scenarios using AI has been well studied and many fleet management companies are adopting the methods to identify the faults in sensors, quality planning, and identifying the adaption of emissions with the age of the vehicle.

3.3 Self Driving and Truck Platooning

Industry 4.0 technologies such as Deep Neural Networks have led to the development of autonomous/self-driving vehicles. As the level of autonomous driving depends on the level of capability of vehicle without human intervention. Level 0 constitutes human-directed vehicles. Level 1 and Level 2 autonomous vehicles provide driver assistance such as lane assistance, and cruise control. Level 3 autonomous vehicles have the environmental detection capability to make informed decisions with human override. Level 4 has a high capability of autonomous driving where human interaction is not required in most scenarios but still requires human intervention. Level 5 autonomous vehicles are fully automated cars that don't require human interaction. Previous works on lane assistance [61–65], pedestrian detection [66–71], vehicle detection [72–74], object detection [75], traffic sign recognition [76–78], self-driving [79–81], determination of turning radius and lateral acceleration in cargo [82] has shown the great success of autonomous cars. Platooning-based video information sharing Internet of Things framework has been proposed to enhance the safety and stability of autonomous vehicles [83]. However autonomous trucks are still a challenge, with the success of autonomous vehicles many companies are focusing on autonomous trucks. Autonomous vehicles are safer than human-controlled vehicles enhancing safety and avoiding human errors especially in preventing collisions during inclement weather and driver behaviors.

The underlying technology for autonomous cars and trucks is similar however trucks need to be able to sense conditions in advance. The current level of autonomy in trucks is within the range of Truck Platooning and is expected to achieve Level 4 by 2024 [84]. Truck Platooning is where tucks travel together connected by a computer and automated driving system. The trailing vehicles adapt and react based on the lead vehicle's action resulting in the semi-autonomous truck. There are three areas of Truck Platooning studies: fuel consumption in truck platoons [85, 86], energy efficiency [87], speed control and control design, communication methods, and interaction for autonomous driving. Various studies have been performed on vehicle platooning such as the

prediction of drag force [88], and the effect of surrounding traffic behavior using machine learning. A neural network structure for deep deterministic policy gradient (DDPG)-based proportional integral derivative (PID) has been developed for vehicular platoon control [89]. This method uses reinforcement learning to find an optimal strategy for deciding based on collision, maintaining relative position, and host responses. The DDPG network consists of an actor network with 1 input layer, 2 hidden layers with 150 and 100 neurons, and an output layer with ReLU activation function, and a critic network with 2 input layers, 3 hidden layers with 150, 200, and 100 neurons and 1 output layer. This method has high interpretability and stability compared to the traditional DRL method achieving a maximum speed error of 0.02-0.08m/s lower than the conventional PID controller and a maximum distance error of 0.77m less than the PID controller.

An CNN and LiDAR-based obstacle detection model with the bird-eye view (BEV) map generation was proposed in [90]. The traditional LiDAR point clouds were used by merging continuous frames and eliminating the ground. Different CNN models namely YOLOv3-tiny, YOLOv3-tiny, XNor, HetConv, and Stride-YOLO are trained on three different LiDAR projection maps (BEV maps of LiDAR point clouds): 1) c1f3g1: H-Map (1 channel) with a combination of three successive frames and elimination of the ground, 2) c3f3g0: HDD-Map (3 channel) with the combination of three successive frames and without elimination of the ground, 3) c3f3g1: H-Map (1 channel) with the combination of three successive frames and elimination of the ground. An Intelligent self-driving truck system consisting of 1) real-world traffic simulation, 2) high fidelity truck that mimics real truck responses, and 3) an intelligent planning module and multi-mode trajectory planner was introduced by Wang et. al [91]. The realistic traffic simulator contains a mapped network, the traffic controller, and the vehicle meta-information. The high-fidelity truck model is implemented using a real truck's kinematics and powertrain system trained with machine learning approaches. The reinforcement learning technique is used for decision-making and trajectory planning. However, this work is mainly focused on highway performance and is a long way to Level 4 autonomy. Self-Driving provides safety in case of reckless driving making quick decisions than humans. However, with higher technology also comes risks such as technical errors and software attacks, and sensor failure issues due to inclement weather. There is a lot of scope for improvement in this area of research as autonomous trucks are still in the initial phase of hitting the roads, using machine learning/computer vision techniques. Autonomous trucks would require more sensors (LIDAR/

RADAR) and more computational power compared to autonomous cars.

3.4 Predictive Maintenance and Onbard Diagnostics

Maintenance is important in fleet management to improve the reliability and uptime of vehicles. The evolution of Maintenance 4.0 has enabled industries to adopt data-driven approaches for maintenance shifting the paradigm from Reactive Maintenance (RM) to Preventive Maintenance (PM) to Predictive Maintenance (PdM) [92]. Predictive Maintenance uses continuous monitoring to determine when maintenance is required. PdM allows the use of historical data, statistical inference methods, and machine learning techniques for the early detection of failures [93]. Several studies have been made in past for predictive maintenance of automobiles using machine learning [94–96]

Predictive Maintenance in the fleet is of high importance to prevent the downtime of a vehicle which could lead to huge losses, especially for delivery trucks and trucks carrying goods. Prytz et. al. [97] used data mining techniques on the logged vehicle data from trucks to perform predictive maintenance of compressor faults. The dataset used was constructed from Volvo's logged vehicle data (LVD), vehicle data administration (VDA), and vehicle service records (VSR) data. The supervised machine learning algorithms KNN, C5.0, and Random Forest were evaluated on the data, and concluded that using logged vehicle data as a solution for predictive maintenance is feasible. However, this study is only concentrated on Volvo's data, but different manufacturers adopt different engines or systems which affect the maintenance of vehicles. The imbalanced datasets pose a big challenge for training machine learning models. The importance of independency for train and test datasets is evaluated by using the Synthetic Minority Over-Sampling Technique (SMOTE) to balance the data. The effect of the prediction horizon (PH) which is defined as the period of interest for the maintenance of classification was evaluated. The author focused on data mining and machine learning techniques. Artificial Neural Networks would be another choice as the data processing required a lot of knowledge regarding the data and data preprocessing. In [98], a parallel stacked autoencoder was used to obtain low dimensional representations from the massive amount of high dimensional logged vehicle data collected from Volvo's heavy-duty trucks, and the embeddings are passed to an autoencoder to predict the Remaining Useful Life (RUL). This

study is very useful in deploying less computationally expensive model on the vehicles. The use of stacked autoencoders improves the performance by 6.31% with 99.7% data reduction and 23.03% with 86.99% data reduction. Scheduling maintenance becomes easy and reduces the downtime if the time between failures (TBF) can be estimated based on historical data. Chen et. al. [99] modeled the maintenance data collected from a fleet company using DCNN to predict the TBF. The maintenance data included parameters such as times of engine experience maintenance, age of the vehicle, cumulative miles when the failure occurred, model of the vehicle, model year, the registration date of the vehicle, type of vehicle, workstation of vehicle, and area of the vehicle. DNN model is trained on the input features and compared the performance with Bayesian regression, k-nearest neighbors, and decision tree algorithms and found that the DNN model has the lowest root mean square error of 366.73 days using historical maintenance data and 363.07 days using historical maintenance data with GIS data that includes rainfall, days of rainfall ¿1mm, max temperature, min temperature, and days of air frost during December and February. This work included processing nominal maintenance data using an autoencoder to obtain low-dimensional and robust data. The low-dimensional data is concatenated with the remaining features of historical data and GIS data is passed to DNN for training. The weights of the neural network are then analyzed to determine the effect of GIS parameters on the output. This is interesting work, but the knowledge of maintenance followed by the company is required. Considering the type of vehicle, fuel type can improve the results as diesel engines require more maintenance compared to alternative fuel engines. Sun et. al. [100] proposed onboard predictive maintenance with machine learning and deep learning models for malfunction prediction and root cause analysis. Their work involved multiple steps where the first step is to identify the data with a high probability of failure collected using the majority voting method and the Diagnostic Trouble Code (DTC) which are compared with multiple machine learning algorithms such as Naïve Bayes classifier, Decision Trees, Support Vector Machines (SVM) and nearest centroid. When an abnormality is detected, the time-series data is recorded, and the final step is to perform sensor data level analysis assuming the sensor data are dependent. A convolutional neural network is trained to reconstruct the expected behavior of the sensor so that the deviation in prediction can be identified as a malfunction. The abnormal sensor combinations are then mapped to root causes. CNN performed well in learning information from time-series data compared to ML algorithms. This technique is less expensive as the data

is recorded only in case of abnormality, but the sensor signal selection was based on the domain knowledge of individuals and is assumed to be not independent. Also, a small deviation in recorded values may not show an impact immediately in the histograms. Rengaswamy et. al. [101] studied the effect of dynamically weighted loss and focal loss in a neural network for prognostic and health management analysis of gas turbine engines and air pressure systems in heavy-duty trucks. Multiple models such as Feed-Forward Neural Network, 1D Convolutional Neural Network, Bidirectional Gated Recurrent Unit, and Long Short-Term Memory were evaluated using the Scania truck dataset and obtained an improved classification using a dynamically weighted loss function. The proposed weighted loss function in this work uses the weight variable (D) given by

$$D(f(x), y) = \begin{cases} \frac{|f(x)-y|}{2}, & \text{if } |f(x) - y| < C. \\ |f(x) - y|, & \text{otherwise.} \end{cases} \tag{3.1}$$

The weighted loss function has statistically significant improvement in all models for remaining useful life (RUL) prediction of gas turbines and anomaly detection in air pressure system of heavy-duty trucks. The learning process using weighted loss function depends on the weight of learning error giving more weightage to focus on the larger error data samples preventing the neural network from biasing the prediction. The freight companies are benefiting from Predictive Maintenance based on historical data. Predictive maintenance helps reduce the downtime of the vehicle, lowers the maintenance cost, ensures safety preventing sudden failures. The existing studies however are limited to predicting remaining useful life (RUL), time between maintenances (TBM), failure in individual components, abnormalities in sensor functioning etc. There are very limited studies on estimating the maintenance cost or identifying the parameters in estimating the maintenance cost. This could give customers an idea to choose a vehicle based on the requirement and maintenance cost.

3.5 Conclusion

The high level of digitization has changed every industry drastically over the past few years. Artificial Intelligence is a cutting-edge technology that decides human intelligence. The potential of AI in the automotive industry mimics human action driving the development of AI in automotive. IoT and cloud technology have enabled the capability to process large volumes of data

paving way for intelligent vehicles. The need for AI in the automotive industry is fueled by the increasing demands for new features, and incorporating new technologies, and the lack of truck drivers has made modest progress since 2017. This resulted in a drastic change in evolution in the past decade. Newer cars come with driverless, intelligent decision-making systems, safety, driver assistance, fuel-efficient, and lower emissions. AI is becoming an essential part of automotive manufacturing, supply chain, and automobiles themselves for self-driving expanding the automotive industry. Companies adopting AI-based technologies and solutions can gain a significant advantage in the coming years. AI, IoT, and Machine Learning are changing the way people think about vehicles extending the features to heavy-duty vehicles. In this literature review, applications of AI in heavy-duty vehicles were introduced by discussing the data analysis/machine learning techniques that can improve fuel efficiency/predict fuel consumption, predict emissions, identify abnormalities in vehicle performance, predictive maintenance, calculate Remaining Useful Life (RUL) and Time Between Maintenances (TBM), truck platooning for self-driving. A review of insightful research efforts has been presented in each of these areas. Industries adopt these technologies based on the research available. There are many research papers and review papers available in the field of automotive, AI in autonomous vehicles, AI in passenger vehicles, AI in predictive maintenance, etc. This comprehensive literature review is focused mainly on applying modern Artificial Intelligence technologies to heavy-duty vehicles as freight transportation is one of the major contributors to climate change, health impacts, pollution, and the country's economy. Even a small improvement in fuel efficiency and lower emissions, preventing downtime will help freight companies save a lot of money and reduce environmental impact. Although electric vehicles, hybrid vehicles, and autonomous vehicles are major trends using computer vision and decision making there is a long way for heavy-duty trucks to reach that point. Availability of data is the major requirement for machine learning/deep learning techniques. The collection of data from different vehicles using different fuel types such as diesel, natural gas, and propane vehicles could help new studies using machine learning in comparing the performance such as fuel consumption, emissions, and maintenance of alternate fuel, diesel vehicles. Reinforcement learning could be the future of automotive that can improve performance by interacting with the environment. More focus is needed on improving the performance of machine learning models as a false positive prediction can cost a lot, especially in the scenarios such as predictive maintenance. One of the major challenges in using

machine learning is the limitation in publicly available datasets which could help in studies related to fuel consumption, emissions, and predictive maintenance of heavy-duty trucks and comparing the performance of different fuel types based on the activity of trucks, especially for vocational trucks.

Chapter 4

Trip-Based Modeling of Fuel Consumption in Heavy-Duty Vehicles Using Artificial Intelligence

Heavy-Duty Trucks contribute about 20% of fuel consumption in the US. The fuel economy in Heavy-Duty Vehicles is affected by several real-world parameters like road parameters, driver behavior, weather conditions, vehicle parameters, etc. Although modern vehicle complies with emission regulations, potential malfunction of the engine, regular wear, and tear or other factors could affect the vehicle's performance. Predicting fuel consumption per trip based on dynamic on-road data can help the automotive industry to reduce the cost and time for on-road testing. Data modeling can easily help in diagnosing the reason behind fuel consumption with the knowledge of input parameters. In this paper, an Artificial Neural Network (ANN) is implemented to model fuel consumption in modern heavy-duty trucks for predicting the total and instantaneous fuel consumption of a trip based on a very few key parameters such as Engine Load (%), Engine Speed (rpm), and Vehicle Speed(km/h). The instantaneous fuel consumption can help predict the pattern in fuel consumption for optimized fleet operations. In this work, the data used for modeling is collected at a frequency of 1Hz during on-road testing of modern Heavy-Duty Vehicles at WVU CAFEE using the Portable Emissions Monitoring System (PEMS). The performance of the Neural Network is evaluated using Mean Absolute Error and Root Mean Square Error. The model is further evaluated on data collected from a vehicle on-road trip. The study shows that Artificial

Neural Networks performed slightly better than other machine learning techniques such as Linear Regression, and Random Forest with high R^2 and lower Root Mean Square Error.

4.1 Introduction

The fuel efficiency of heavy-duty trucks can be beneficial not only for the automotive and transportation industry but also for the country's economy and global environment [102, 103]. The cost of fuel consumed contributes about 30% of heavy-duty truck life cycle cost. Reduction in fuel consumption by just a few percent can save a lot of costs for the transportation industry [104, 105]. Effective and accurate estimation of fuel consumption (fuel consumed in L/km) can help analyze emissions as well as prevent fuel-related fraud. As per EPA reports, 28% of total greenhouse gas emissions come from transportation (heavy-duty vehicles and passenger cars) [106]. The United States Environmental Protection Agency (US EPA) has introduced Corporate Average Fuel Economy (CAFÉ) standards enforcing automotive manufacturers to be compliant with the standards to regulate fuel consumption [107, 108].

Several studies have been presented in the past for evaluating the fuel efficiency of vehicles using simulation-based models, and data-driven models. A simulation model was developed based on engine capacity, fuel injection, fuel specification, aerodynamic drag, grade resistance, rolling resistance, and atmospheric conditions with simulated dynamic driving conditions to predict fuel consumption [109]. A statistical model which is fast and simple compared to the physical load-based approach was developed to predict vehicle emissions and fuel consumption [110]. The impact of road infrastructure and traffic conditions [111–113], driver's behavior [114,115], weather conditions [116, 117], and the ambient temperature on fuel consumption were studied and learned that fuel consumption can be reduced by 10% with eco-driving influences. The era of big data and artificial intelligence has enabled the modeling of huge volumes of data for companies to reduce emissions and fuel consumption. Machine learning techniques such as Support Vector Machine (SVM) [118], Random Forest (RF) [119], and Artificial Neural Networks (ANN) [120, 121] are widely applied to turn data into meaningful insights and solve complex problems. These techniques have been applied to estimate emissions and fuel consumption in motor vehicles [29], trucks [36], ships [122], and aircraft [123]. The comparison of previous studies has been shown in Table 4.1

Table 4.1: Previous studies related to Fuel Consumption

Predicted Variable	**Input Parameters**	**Method**	**Result**	**Reference**
Fuel consumption in passenger cars	Cubic Capacity, Quantity of cylinders, Quantity of valves, Maximum Power, Maximum Torque, Compression Rate, Kerb Weight of Vehicle, Type of Engine, Fuel Injection, Type of charge, Gearbox, Drivetrain	ANN (22-10-3) ANN (20-10-3)	$R^2 \geq 0.98$; RMSE = 5-8; $R^2 \geq 0.98$; RMSE = 6-10	[29]
Fuel consumption in trucks	Road Gradient, Torque % at the start, Torque % at the end, Average Acceleration Gross Vehicle Weight, Road curvature radius, Longitudinal Profile , Variance at 3m, Longitudinal Profile Variance at 10m, Longitudinal Profile Variance at 30m, Vehicle Speed, Cruise control, Used Gear	SVM RF ANN	$R^2 = 0.83$; RMSE = 5.12; MAE = 3.56 $R^2 = 0.87$; RMSE = 4.64; MAE = 3.21 $R^2 = 0.85$; RMSE = 4.88; MAE = 3.46	[36]

Predicted Variable	**Input Parameters**	**Method**	**Result**	**Reference**
Fuel consumption based on floating vehicle data	Driver Gender, Driver Age, Transmission Type, Fuel Type, Weight, Mileage, Speed, Time, Location	ANN(9-4-1) ANN(9-6-1) ANN(9-8-1) ANN(9-10-1) ANN(9-12-1)	MSE = 0.00032692 MSE = 0.00037202 MSE = 0.00019523 MSE = 0.00009996 MSE = 0.00025849	[33]
Brake specific fuel consumption and exhaust temperature in diesel engine	Engine Speed, Brake Mean Effective Pressure, Injection Time,	ANN(3-7-2)	Mean Relative Error for BSFC=1.93% Mean Relative Error for Exhaust Temperature = 2.36	[34]
Fuel Consumption	Engine Size, Distance, Fuel Type, Speed, Weight.	RBFNN(5-15-1)	Maximum Error Percentage = 0.024 Absolute Avg. Error = 0.022071	[124]

While the current approaches determine the fuel consumption of the vehicle, combining these techniques with data help in identifying the parameters that cause the anomaly as the malfunction can be due to wear and tear of the engine, improper maintenance, engine failure, exhaust after-treatment system, external factors like climate, traffic, road conditions, etc. Most of the studies in the literature were limited to passenger cars, light-duty vehicles, and heavy-duty vehicles, or based on a huge number of parameters or limited dynamic data collected during on-road trips. However, modeling modern heavy-duty trucks with very few parameters is much more complicated. This current study models fuel consumption in modern heavy-duty trucks based on Portable Emissions Monitoring System (PEMS) data collected during on-road testing. An artificial neural network is developed to predict the total fuel consumed by a vehicle in a trip based on very few key parameters such as engine load (%), engine speed (rpm), and vehicle speed(km/h). The model also gives the trend in fuel consumption for the trip which give insights into the diagnostic performance of truck affected by the input parameters. The model can predict the total fuel consumed more accurately with a mean absolute error of 0.0014 and root mean square error of 0.0025 compared to other techniques such as linear regression [125] and random forest [119].

4.2 Methodology

Regression analysis is performed using Machine Learning techniques to estimate the fuel consumption of modern heavy-duty trucks using PEMS data. The preprocessed dataset related to a single vehicle containing 672658 rows of actual torque (ft-lb), vehicle speed (km/h), and engine speed (rpm) is used as input for models. Figure 4.1 shows the overall workflow for this work.

4.2.1 Data Collection and Pre-Processing

Data collection methods such as onboard emission measurement [126], laboratory measurement and tunnel study [127] have been used in the past. An on-road data collection method using PEMS is increasing nowadays making it possible to collect real-world fuel consumption and emission data [128] and proved to be reliable [129]. A PEMS device is mounted on a vehicle to collect data at 1Hz frequency during on-road testing. PEMS software outputs the data second by second into a CSV file for each trip. Over 100 parameters such as fuel rate (L/h), engine speed (rpm), speed (km/h),

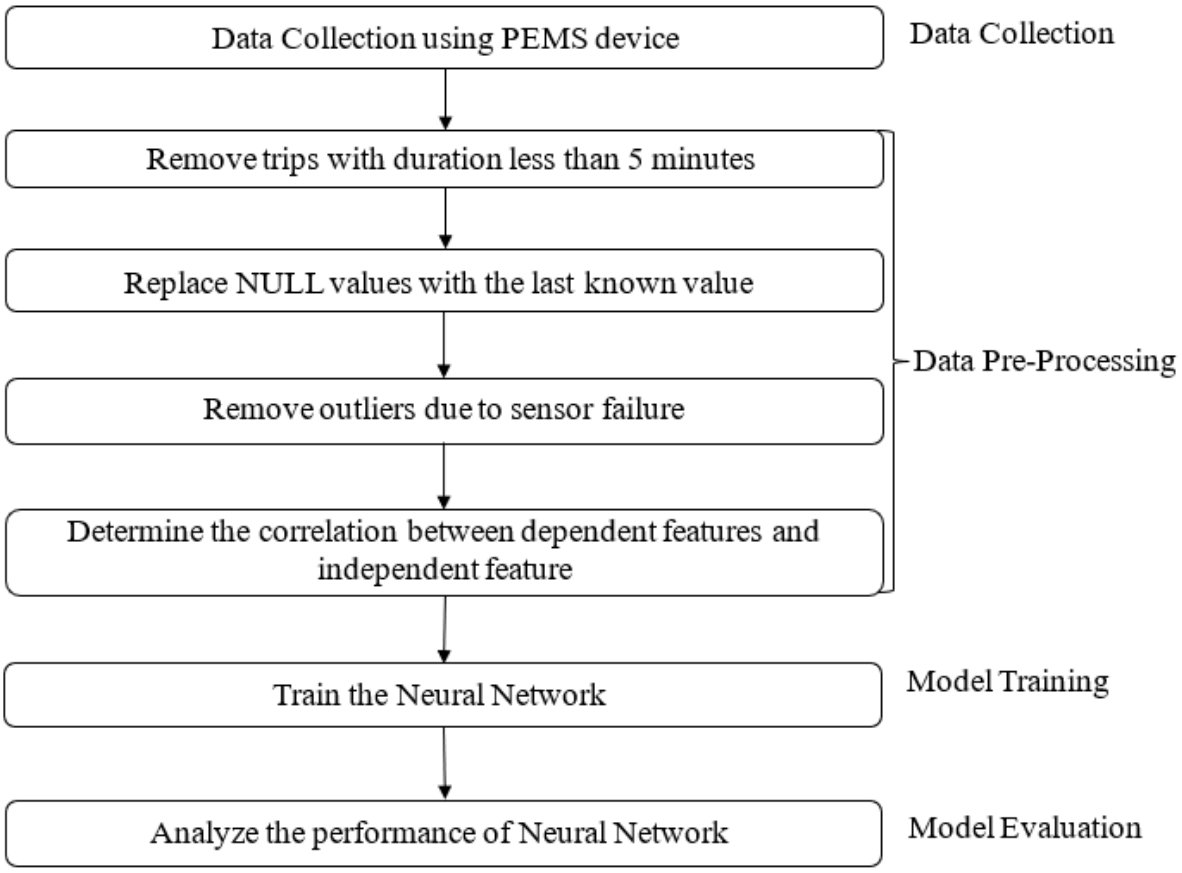

Figure 4.1: Workflow for fuel consumption modeling

gas temperature, CO2, NOx, GPS altitude, GPS longitude, GPS latitude, etc. were collected for each trip based on data logger settings. Data is collected from two heavy-duty trucks with the same make/model of the engine by Detroit diesel manufactured in 2016. The vehicles are tested on multiple days with different routes, drivers, and conditions. However, modeling with too many parameters can cause the neural network model to overfit resulting in poor performance. Hence, a subset of features is selected based on data pre-processing, previous studies, and domain knowledge which are shown in Table 4.2.

Feature Engineering

For better modeling, the data collected must be representative. The raw dataset contains noise/missing values, redundant values, and outliers due to failure in the sensor or sensor not enabled for recording. With feature engineering, the raw data is transformed into features that better represent the relation between features to the predictive model, resulting in better performance

Table 4.2: A subset of features selected for fuel consumption modeling

Variable Name	**Description (unit)**
Trip Number	Index for each trip file
Engine Speed (rpm)	Speed of engine (rpm)
Trip Distance (km)	Distance traveled for a trip (km)
Vehicle Speed (km/h)	Speed of vehicle at the current instance (km/h)
Fuel Temperature (degC)	The temperature of the fuel (oC)
Fuel Rate (L/sec)	Amount of fuel consumed by a vehicle at that point (L/sec)
Accelerator Pedal Position (%)	Position of the accelerator pedal (%)
Actual Torque (ft-lb)	Torque of engine (%)
Power (bhp)	Rated power of the engine (bhp)
Engine Load (%)	Load on the engine (%)

accuracy. Since the data is collected from multiple sensors, noise or unwanted values can cause a problem in modeling with machine learning as the model tries to interpret noise as a pattern in data. This work makes a few assumptions to eliminate noise and missed values.

1. The first 5 minutes (300 seconds) of the data have been removed as the cold start of the engine records default values for most of the sensors.

2. A failed/disabled sensor logs no data, hence columns of data with 80% of missing values are eliminated.

3. An intermittent missed value is imputed with the most recent value logged or a zero.

Once the missing values are dealt with, the handling of outliers that skew the distribution of data based on domain knowledge is implemented. In this study, the outliers are identified by calculating the Z-Score, and a statistical approach of the interquartile range is used to plot the outliers (Figure 4.2). Not all the values identified with the statistical approach may be outliers, hence it is important to analyze the outliers before removing them. In some cases, the outliers might indicate essential conditions that need to be considered for analysis. For example, a failure in the sensor might log a default value for the feature. Such edge conditions are analyzed carefully before removing the data point from modeling.

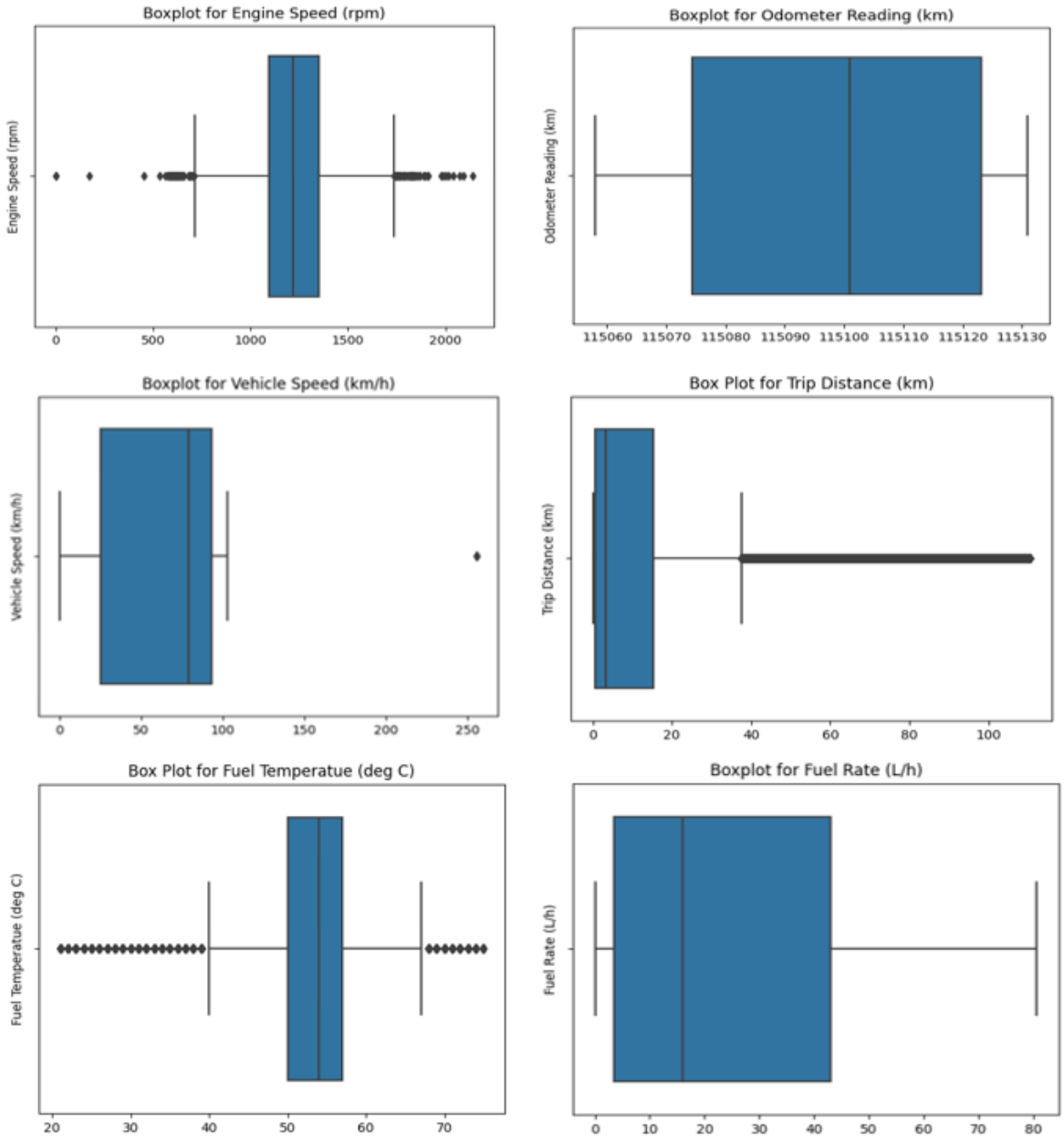

Figure 4.2: Boxplot for Outlier Detection

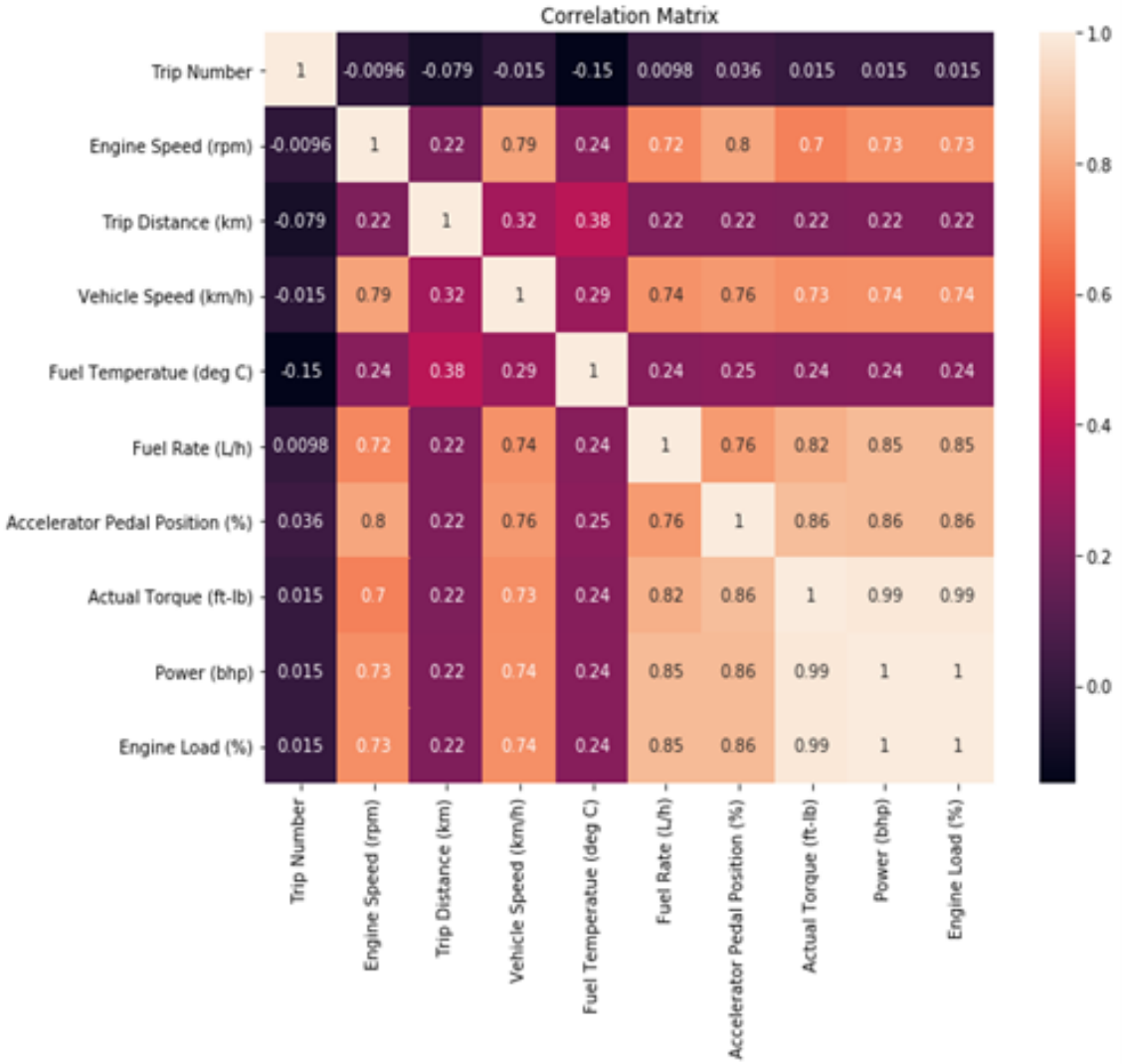

Figure 4.3: Feature Correlation Matrix

4.2.2 Feature Association

The interpretation of the regression model is complex when the independent variables are multi-collinear. Highly correlated independent variables overfit the model as the change in one variable causes significant change to another. Hence to identify the multi-collinear variables, a correlation matrix that determines the correlation coefficient of each variable with every other variable in the data is shown in Figure 4.3.

The last four rows and columns of the correlation matrix indicate that independent features accelerator pedal position (%), actual torque (ft-lb), power (bhp), and engine load (%) are highly correlated to each other with a correlation coefficient of 0.85 and higher. Hence, to prevent over-fitting of the model only engine load (%) of the four parameters is used in modeling. Feature dimension can further be reduced by identifying the highly correlated features with the target variable fuel rate (L/sec). The recursive Feature Elimination (RFE) and Ridge Regression methods are used to identify and plot the feature importance scores. Feature importance of remaining features engine load (%), accelerator pedal position (%), fuel temperature (deg C), vehicle speed (km/h), trip distance (km), and engine speed (rpm) concerning fuel rate (L/sec) was determined with RFE (Figure 4.4) and Ridge Regression (Figure 4.5) technique and the top 3 features with the highest score are selected for modeling. Based on the feature analysis, three independent features namely engine load (%), vehicle speed (km/h), and engine speed (rpm) with high importance are selected for modeling the dependent feature fuel rate (L/sec) and to identify the pattern in fuel consumption.

4.2.3 Artificial Neural Network

Artificial Neural Network (ANN) is a machine learning technique inspired by biological neurons. ANN consists of multiple neurons which are computational and the connections between neurons determine the functionality of the network [130]. The building block for a neural network is a neuron that represents the weighted sum of inputs passed through a non-linear activation function. Multi-Layer Perceptron (MLP) network is a type of neural network that consists of the input layer, one or more hidden layers, and an output layer. ANN has gained popularity due to its adaptive learning ability and approximating non-linear functions to make predictions [131]. In this study, a feed-forward neural network [132] with a backpropagation algorithm and ReLU [133] activation

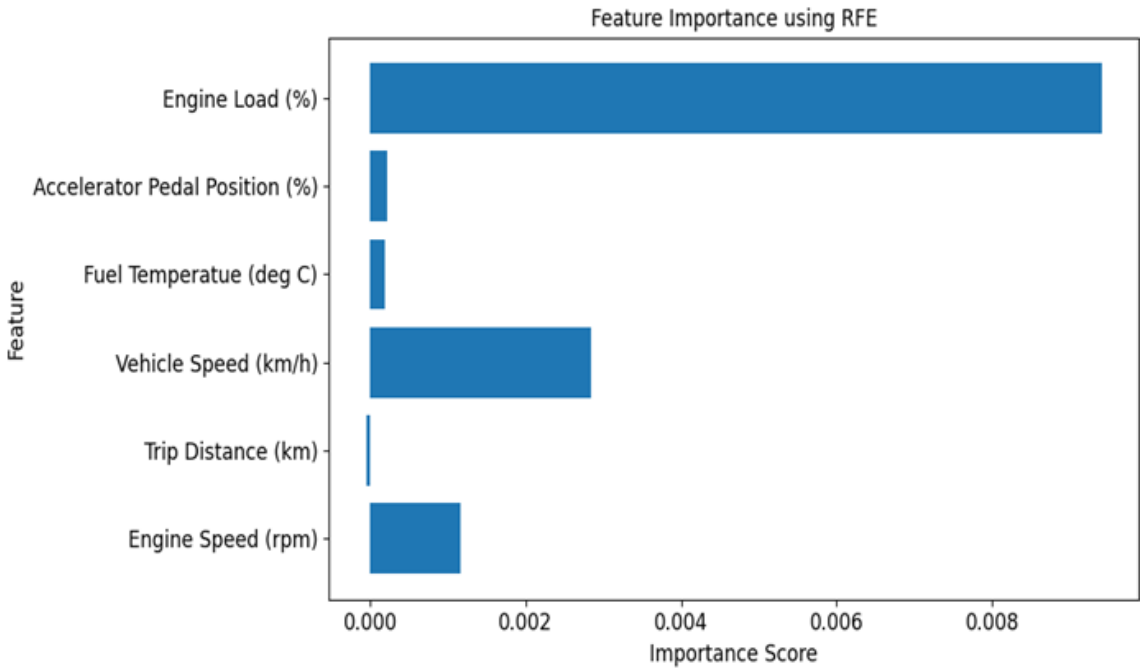

Figure 4.4: Feature Importance using Recursive Feature Elimination

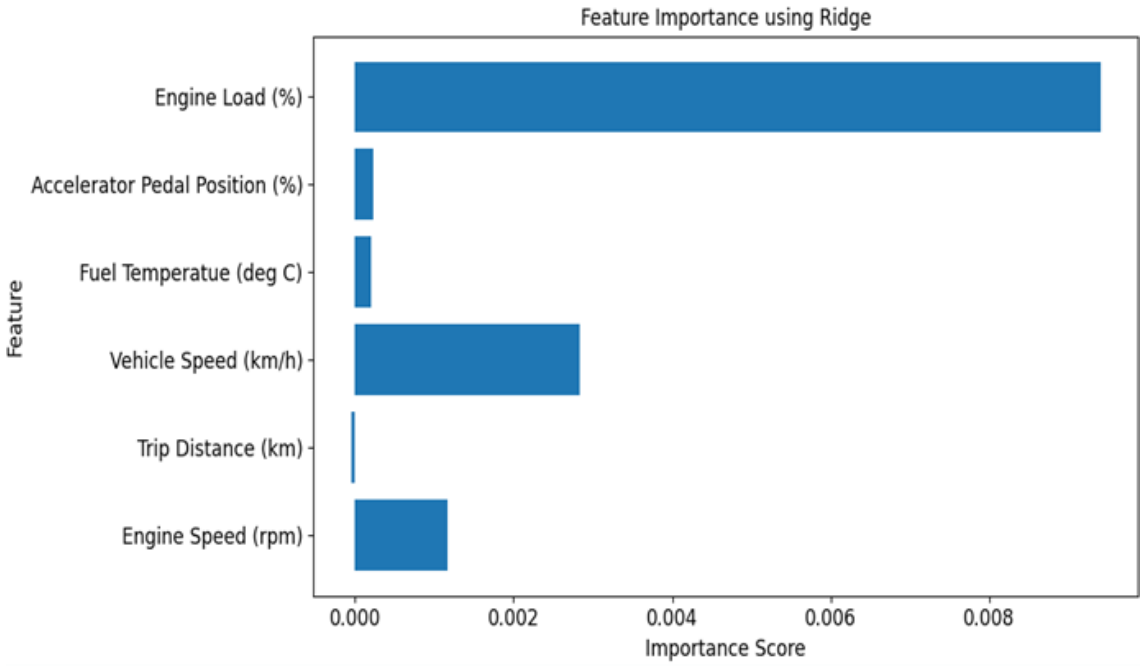

Figure 4.5: Feature Importance using Ridge Regression

is used. The backpropagation algorithm is a learning method to train neurons in a neural network by repeatedly adjusting the weights to minimize the error of prediction. The network used for this work has three inputs to the input layers, two hidden layers with six and eight neurons in the respective layers, and an output layer with a single neuron as shown in Figure 4.6.

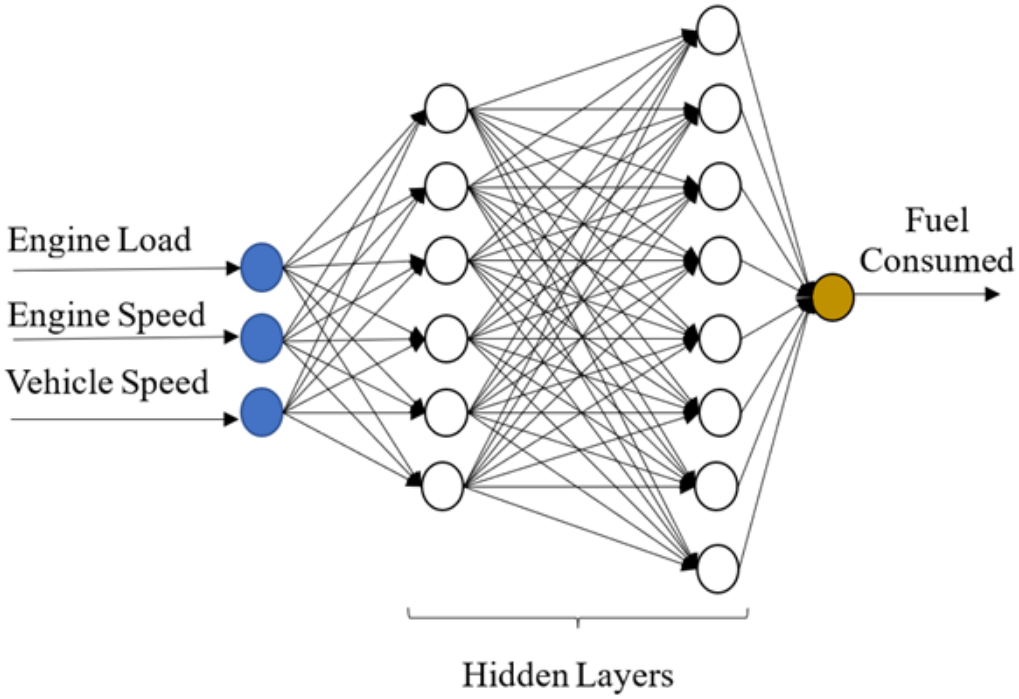

Figure 4.6: Artificial Neural Network Structure

The available dataset of vehicle 1 is divided into train and test sets with 70% to train the network and 30% to test the generalization of the network. The trained model weights are then used to make predictions on unseen test data (a single trip from Vehicle 2). The performance of a neural network depends on many hyper-parameters like the learning rate, number of epochs for training, initial weights, number of hidden layers, and number of neurons in hidden layers. Multiple experiments are performed with different hyper-parameters and the best results for the optimal network are presented in the results section.

4.2.4 Multiple Linear Regression

Multiple Linear Regression (MLR) [125] is the most well-known regression technique where the data is fitted to a straight line to predict output by minimizing a cost function or error. In this study, a multi-variable linear equation given by equation 4.1 is used due to multiple input parameters.

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 \tag{4.1}$$

Where, y is the output and x_1, x_2, x_3 are the input variables with $\theta_0, \theta_1, \theta_2, \theta_3$ being parameters to learn.

4.2.5 Random Forest

Random Forest (RF) [119] is an ensemble machine learning method for regression and classification tasks. This method uses many decision trees, and the outcome is based on predictions of these decision trees. Thus, the accuracy of the model can be improved by increasing the number of trees making it robust to outliers. In this study, the random forest is trained with 100 trees as the performance with more than 100 trees did not improve much and is computationally expensive.

4.3 Performance Measures

The performance of the machine learning model for the regression problem is evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2).

4.3.1 Mean Absolute Error

Mean Absolute Error (MAE) is the measure of error between the predicted value and the actual value given by equation 4.2.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |x_i - \hat{x_i}|, \tag{4.2}$$

where x_i is the measured fuel consumption and $\hat{x_i}$ is the measured fuel consumption at the same instant of time, and N is the number of data points.

4.3.2 Root Mean Square Error

Root Mean Squared Error (RMSE) is the square root of the average squared difference between the predicted value and the actual value given by equation 4.3. The smaller the value, the closer the predicted values are to actual values.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x_i})^2}, \tag{4.3}$$

where x_i is the measured fuel consumption and $\hat{x_i}$ is the predicted value at the same instant of time, and N is the number of data points.

4.3.3 R-Squared

R-Squared or coefficient of determination (R^2) is the statistical measure of variance for the dependent variable explained by the regression model given by equation 4.4.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(x_i - \hat{x_i})^2}{\sum_{i=1}^{N}(x_i - \overline{x})^2}, \tag{4.4}$$

where x_i is the actual measured fuel consumption at time i, $\hat{x_i}$ is the predicted value at time i, $\overline{x}$ is the mean of data given by $\frac{\sum_{i=1}^{N} x_i}{N}$, N is the number of data points.

4.4 Results and Discussion

This study presents the fuel consumption modeling in modern heavy-duty vehicles using PEMS data under various driving conditions, different routes, and external factors. Engine Load (%), Engine Speed (rpm), and Vehicle Speed (km/h) are used as input for the ANN. The neural network is trained with hyper-parameters such as a batch size of 1024, and a learning rate of 0.0001 based on hyper-parameter tuning results for 100 iterations. During each epoch, the loss for each data item/batch in the training dataset and validation dataset is calculated. The loss plots shown in Figure 4.7 indicate the mean absolute error (MAE) and mean square error (MSE) on both training data and validation data.

The minimum generalization gap of training and validation data loss plots indicates a good fit. The generalization of the neural network is tested on test data collected from a single trip of another vehicle. From Figure 4.8, the data points close to the line indicate the neural network model can

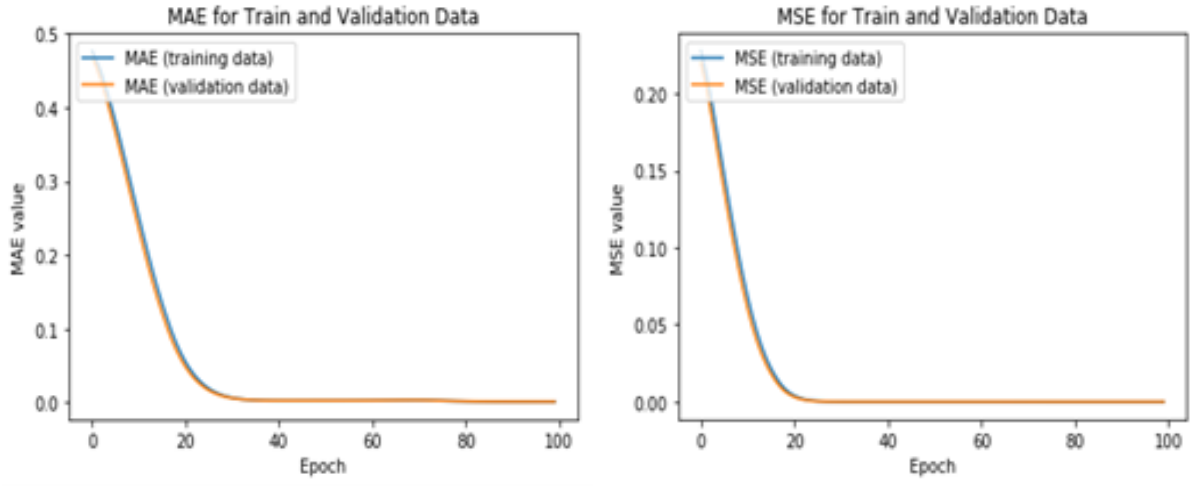

Figure 4.7: Neural Network Performance plots. Mean Absolute Error (Left Panel) and Mean Square Error (Right Panel) plots during ANN training for training and validation data.

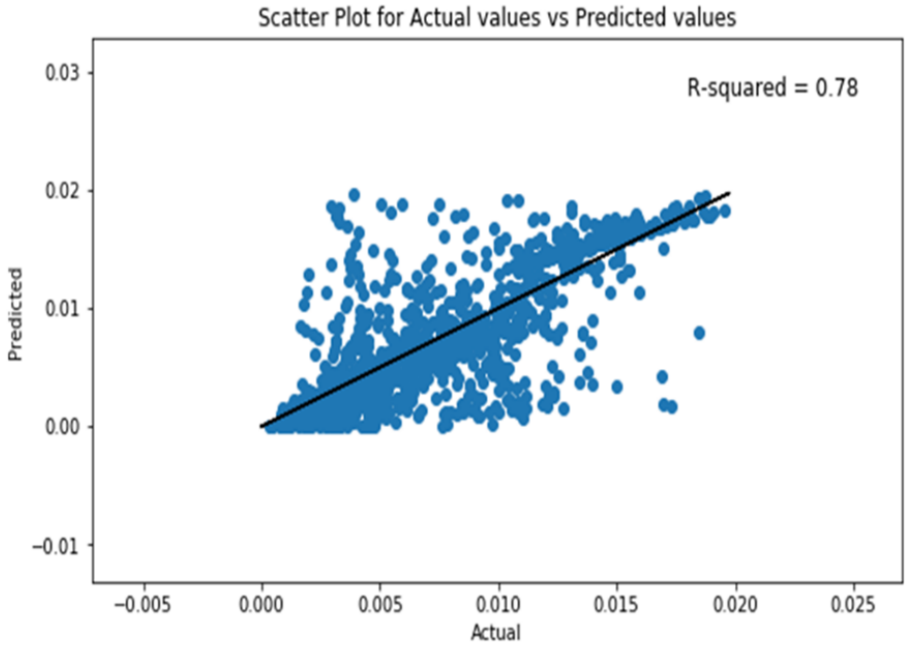

Figure 4.8: Scatter plot for the predicted data and the actual measured data

Table 4.3: Comparison of performance on the test dataset

Model	**MAE**	**RMSE**	R^2	**Fuel (L) Actual**	**Fuel (L) Predicted**	**% Error**
Artificial Neural Network	0.0014	0.0025	0.7806	10.0589	10.0558	0.03%
Linear Regression	0.0018	0.0029	0.7340	10.0589	10.05720	-5.09%
Random Forest	0.0018	0.0030	0.7210	10.0589	9.5589	4.96%

accurately predict fuel consumption with few errors. The points far away from the regression line indicate outliers in data due to sudden transitions in vehicle speed and engine speed which the neural network could not capture.

To determine the total fuel consumed by a vehicle, the cumulative fuel consumption is calculated by adding the instantaneous fuel rate values every second. The performance measures described in section 3 are used to evaluate the model and the values obtained are MAE: 0.0009L, and RMSE: 0.0021L for the training data. The R^2 value of 0.7806 on the test data and 0.7762 on the train data indicate that the neural network model is generalized well for unseen data.

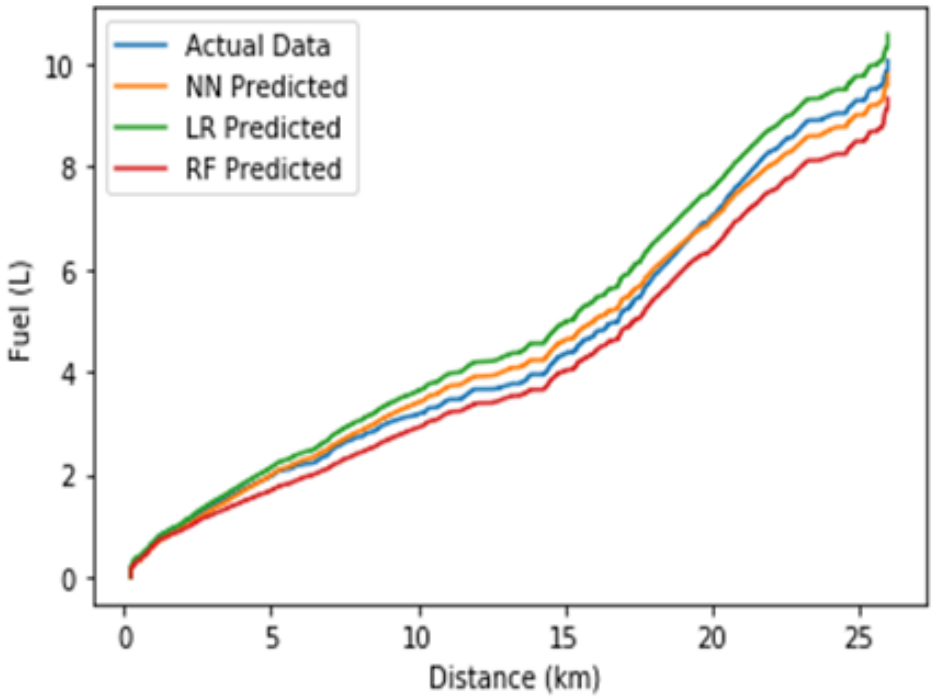

Figure 4.9: Comparison of predicted total fuel consumed per trip using different machine learning algorithms

Table 4.4: Comparison of performance on the test dataset

Vehicle Make/Model	**Engine Year**	**MAE**	**RMSE**	R^2	**Actual Fuel Consumed (L)**	**Predicted Fuel Consumed (L)**
Cummins X15 450	2017	0.0010	0.0015	0.8823	10.67	9.98
Navistar A26 A400MT	2017	0.0012	0.0015	0.7877	5.51	5.57
Paccar MX-13	2016	0.0005	0.0008	0.8005	2.00	2.09

Table 4.5: Performance on the test dataset using the fine-tuned model

Model	**MAE (train)**	**RMSE (train)**	R^2 **(train)**	**MAE (test)**	**RMSE (test)**	R^2 **(test)**	**Actual Fuel (L)**	**Predicted Fuel (L)**
ANN	0.0006	0.0009	0.9723	0.0006	0.0010	0.9443	10.77	10.44

4.4.1 Comparison of results

The cumulative fuel consumption with distance is plotted against the actual data to determine how well the neural network has predicted the total fuel consumption. To evaluate the performance, the neural network predictions are compared with predictions of linear regression, and random forest. The performance metrics MAE, RMSE, and R2 are compared in Table 4.3. Figure 4.9 shows the plots for comparison of cumulative fuel consumed for distance traveled for all models. Neural Network prediction is closer to the actual measured data compared to the linear regression model overestimated and the random forest underestimating the cumulative fuel consumed.

Based on the input features that were modeled it is easy to determine the parameter affecting the fuel consumption in case of anomaly. This study presents an efficient and practical method of estimating fuel consumption per trip based on very few parameters for which data is easily available. The cost incurred in modeling the data is very low compared with other simulation methods which also consume more time.

The test results for different vehicles with the neural network model trained on Detroit Diesel, DD13 manufactured in 2016 are presented in Table 4.4

The model is further evaluated on test data from different make and model vehicles and the performance is shown in Figures 4.10, 4.11, and 4.12

The model is later fine-tuned with new data from goods movement trucks with Cummins engine model ZX15 450 manufactured in 2018. During fine-tuning, the model with the pre-trained weights

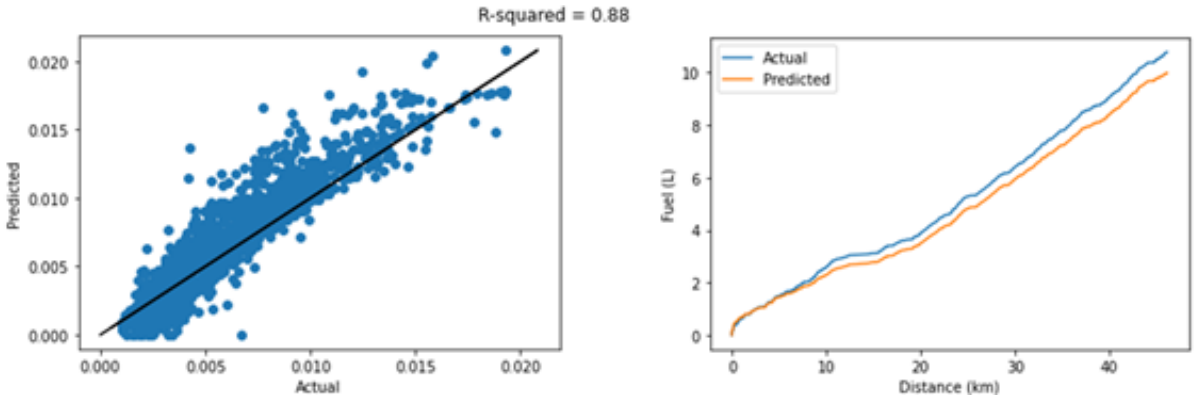

Figure 4.10: Comparison of actual measured values and predicted values for Cummins X15 450 test vehicle (Left Panel). Total fuel consumed vs Distance (Right Panel)

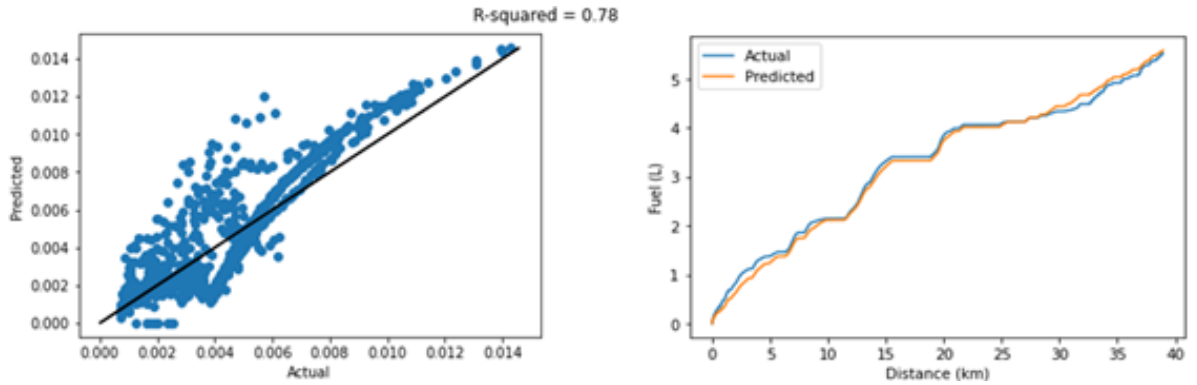

Figure 4.11: Comparison of actual measured values and predicted values for Navistar A26 A400MT test vehicle (Left Panel). Total fuel consumed vs Distance (Right Panel)

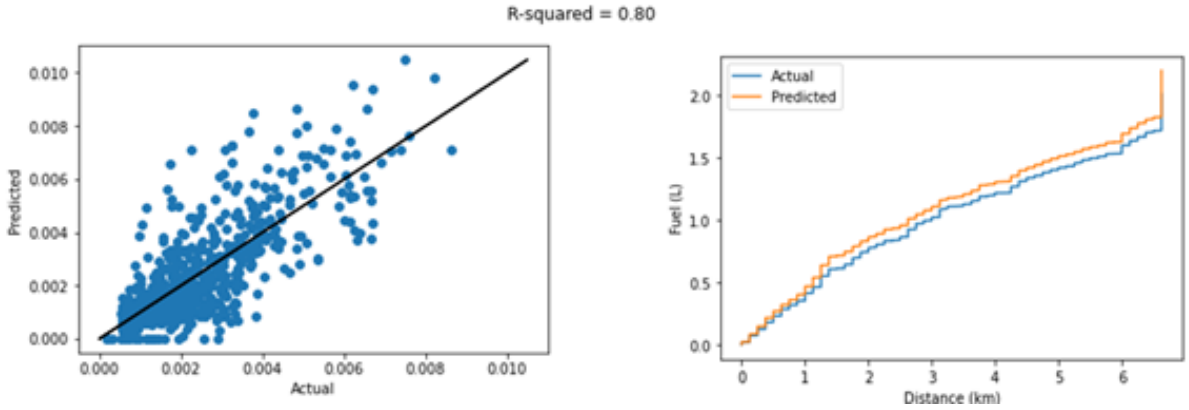

Figure 4.12: Comparison of actual measured values and predicted values for Paccar MX-13 test vehicle (Left Panel). Total fuel consumed vs Distance (Right Panel)

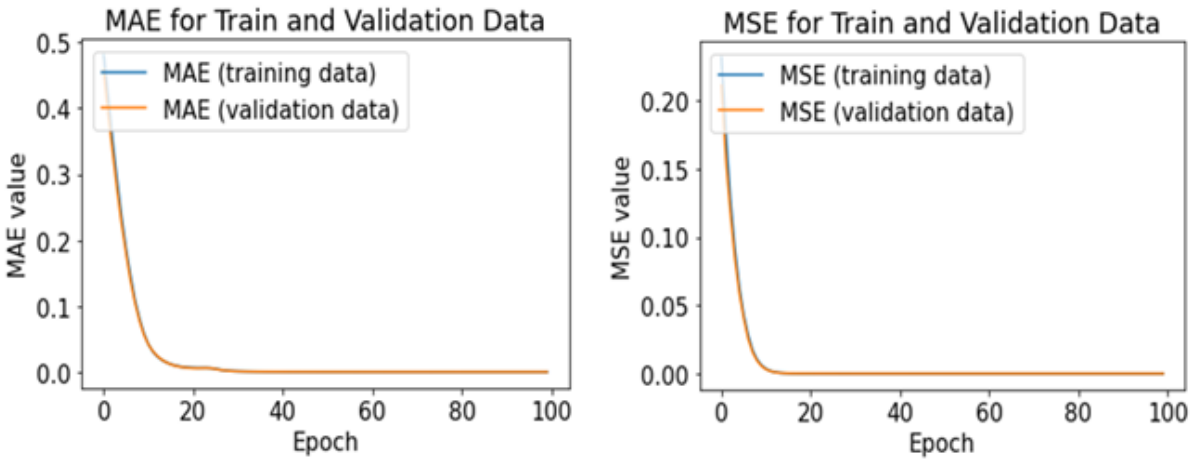

Figure 4.13: Neural Network Performance Plots for fine-tuning. Mean Absolute Error (Left Panel) and Mean Square Error (Right Panel) on training and validation data

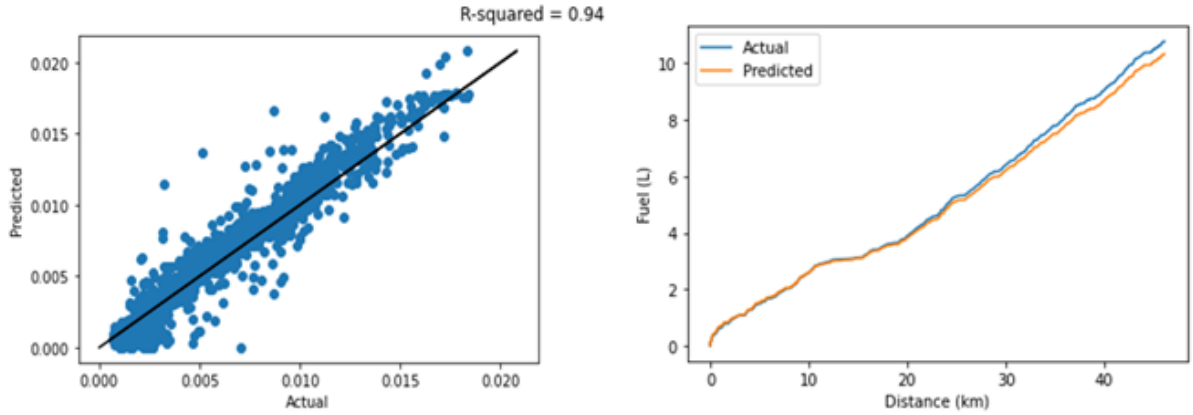

Figure 4.14: Comparison of actual measured values and predicted values for Cummins ZX15 450 test vehicle (Left Panel). Total fuel consumed vs Distance (Right Panel)

is fine-tuned to capture the variation in new data. The fine-tuned model is tested on the same make and model truck manufactured in 2017 and the results are shown in Figure 4.13, Figure 4.14, and Table 4.5. The difference in the trend of fuel rate for different trips indicates how vehicles are operated or driven.

4.5 Conclusion

In conclusion, the study demonstrates the modeling of fuel consumption in modern heavy-duty vehicles using an artificial neural network. An attempt was made to develop a model using very few parameters collected under different conditions. Data from modern heavy-duty trucks with the same make and model, driven by different persons on various routes under different external conditions were used for training the artificial neural network. The model relies on very few parameters that could be obtained easily from a vehicle during the trip. The performance measures MAE, RMSE, and R^2 indicate that accurate prediction can be obtained with the model. The data modeling can help to identify the trend in instantaneous fuel consumption and to calculate the total fuel consumed by the vehicle in each trip which can further help in diagnosing vehicle performance in case of abnormalities. Models that are accurate and fast to predict in real-time will enable the optimization of fuel consumption. The deviation in the predicted values helps identify cases where fuel consumption is affected due to conditions such as tire pressure, driving behavior, etc. This work can further be extended to include other factors such as time, traffic information, road information, GPS data, etc., affecting fuel consumption. As shown in the results sections, the model can be easily fine-tuned to vehicles using different fuel types or vehicles with different make and models. The model can be further integrated into a framework where it can be used to suggest the estimated fuel consumption the parameters such as vehicle speed, engine speed, and engine load.

Chapter 5

Machine Learning for Modeling Maintenance and Repair Costs in Delivery Trucks Using Diesel and Natural Gas Fuels

Fleet management and production companies are constantly trying to avoid maintenance downtime and reduce maintenance costs [134] by scheduling maintenance needs ahead of failure. Since the maintenance costs can represent about 15-60% of the cost of produced goods depending on a specific industry, Fleet Management Systems are adopting better maintenance planning to improve the reliability of transportation that require less unplanned maintenance due to the demand for more uptime.

Heavy- and medium-duty trucks contribute about 26% of greenhouse gas emissions [135] resulting in the emission standards to be set forth. To comply with the standards diesel engines are equipped with a complex after-treatment system, and this requires more frequent and more intense maintenance. In recent years the automotive industry turned to alternative fuels such as natural gas and propane fostering the development of natural gas and propane powertrain systems for heavy- and medium-duty vehicles offering soot-free combustion. It also supported the development of electric powertrain systems known for their zero emissions. Despite the efforts and need for reducing greenhouse emissions, alternative fuel vehicles (AFVs) haven't increased significantly in global transportation [136]. One of the main reasons for fleet companies not adopting AFVs even

though they are better suitable for the task is the lack of knowledge on vehicle maintenance/repairs and its associated cost. Moreover, the region of operation, the type of vehicle operation (delivery trucks, refuse trucks, school buses, goods movement, etc.), and seasonal temperature changes also affect the duty cycle which impacts the maintenance and repair costs.

To fill the gap, we leverage the data collected by WVU CAFEE and a partnership from an ongoing study that has collected vast quantities of vocation-specific vehicle activity data. The partnership with fleet companies allows for additional information related to maintenance costs and fleet-specific maintenance practices of alternative fuel vehicles. In this work, a detailed analysis of the maintenance for delivery trucks using diesel and natural gas fuels is performed. Machine learning techniques are used for predicting the maintenance cost (MC) per mile. The research conducted here helps fleet management companies to understand the advantages and disadvantages of alternative fuel vehicles over heavy and medium-duty vehicles and make informed procurement decisions.

5.1 Introduction

Heavy-duty vehicles primarily operated by diesel fuel are used worldwide for various activities such as good transportation, delivery, refuse trucks, etc, and have a significant impact on the environment due to high emissions [137]. Hence a significant push for switching to cleaner fossil fuels such as natural gas with relatively low emissions [138] and comparison with diesel fuel has been studied [139]. Natural gas/alternative fuel vehicles are, however, not prominent due to their high price, less infrastructure availability, and unknown total cost of ownership. Understanding the maintenance requirements and cost involved in maintenance and repairs (MR) for natural gas engines and the comparison with diesel vehicles is important in promoting alternative fuel vehicles.

In recent years, the high computational power, big data, and data storage enabled companies to derive meaningful insights from huge volumes of real-time noisy data, especially for tasks such as predictive maintenance (PdM) and remaining useful life (RUL) using machine learning techniques. Determining the average time until the next maintenance based on the current state of the automobile/engine can be useful in scheduling maintenance and lowering maintenance costs. Predictive maintenance in automobiles using machine learning requires modeling huge volumes of

sensor data, hence a representation learning to convert high dimensional data to low dimensional and predicting vehicle faults was proposed in [98] using parallel stacked autoencoder. Machine Learning techniques such as support vector machines (SVM) [140], random forest (RF) [119], feed-forward neural networks (NN) [141], and Gaussian processes (GP) [142] are evaluated for predictive maintenance in automotive engine components using simulated data [143]. Several deep learning (DL) algorithms have been used in identifying faults and estimating remaining useful life (RUL) in various automobile fields due to their ability to extract features automatically. An ensemble approach combining RUL estimation from a similarity-based curve matching technique and bidirectional recurrent neural network (RNN) was proposed in [144]. A multiscale convolutional neural network was introduced for bearing RUL estimation using the bilinear interpolation of time-frequency representations as input to the deep learning model [145]. Prognosis involves the effect of time, hence as the time step increases, the accuracy of the prognosis decreases. Hence a sequence-based recurrent neural network (RNN) model using the vibration signals of defect rolling bearings was analyzed in [146] with more accurate prediction than the incremental training. Long Short-Term Memory (LSTM) Neural Networks and Random Survival Forest (RSF) [147] are applied for predicting the component failure probabilities in lead-acid batteries of heavy-duty vehicles using few data collected during workshop visits [148] and LSTM performed significantly better compared to other techniques. Multi-sensor fault detection, fault identification, isolation, and health index forecasting were performed using deep convolutional neural networks in autonomous vehicles [149]. The remaining fatigue life based on health monitoring of automotive suspension was estimated in test cars using LSTM [150].

There are a lot of studies related to maintenance prediction, but studies related to maintenance cost estimation are very limited. To determine the cost involved in corrective maintenance of replacing the failed part has been studied using the Poisson distribution stochastic model [151]. This study has been performed on urban rail vehicles with 45 vehicles used for 5 years. An artificial neural network (ANN) model for classifying the faults and determining the frequency of failure was analyzed [152]. The model includes six vehicle usage parameters such as fuel cost, fuel volume in liters, car mileage in km, normalized fuel cost, normalized fuel volume, and normalized mileage as input parameters with maintenance cost as output. The operating costs were studied for personal vehicles and large commercial trucks based on consumer guides and a review of several sources

of trucking costs respectively based on stop-start conditions, pavement roughness, and inflation [153]. However, the existing studies are limited to estimating maintenance costs for passenger cars or a specific type of maintenance cost in vehicles. There are very few studies related to comparing maintenance in alternative fuel vehicles and estimating the average cost per mile as the maintenance of vehicles is highly impacted by the fuel type and the vocation.

There is a large gap in adopting alternative fuel vehicles due to the uncertainty and unavailability of data and information [136] even though they produce fewer emissions [154]. Hence to fill the gap, this study addresses the following:

1. Lack of real-time and real-world maintenance data records from alternative and diesel-fueled heavy-duty trucks by collecting the data from fleet management companies.

2. Application of the machine learning models on the large volume of maintenance data collected from delivery trucks using natural gas and diesel to understand the maintenance cost patterns.

5.1.1 Gap in Knowledge

A Survey of the literature reveals a large gap in information related to the changes in MR cost as a function of vehicle age. The effect of age on MR cost can be highly variable between fleets since the effectiveness of preventive maintenance strategies lowers the slope of the MR cost as a function of age. Furthermore, the uncertainty and unavailability of this information about AFV contribute to a critical barrier to AFV adoption.

Most of the previous work done in estimating MC of alternative fuel technology has used OEM-prescribed maintenance intervals assumed parts failure rates, and generic periodic and preventive maintenance schedules. However, there has been no study that has involved large-scale fleet participation to document MC from both AFVs and diesel vehicles. Moreover, not all maintenance cost studies are equipped with detailed vehicle telemetry data that characterize vehicle operation. Ultimately, the type of vehicle operation defines the type of maintenance, frequency of maintenance, and the costs associated. The proposed study presents an innovative platform to collect both detailed maintenance records from AFVs and modern diesel fleets as well as link them to vehicle telemetry data. The study would positively impact the alternative fuel infrastructure stakeholders and AFV manufacturers by helping them identify the barriers associated with AFV adoption in

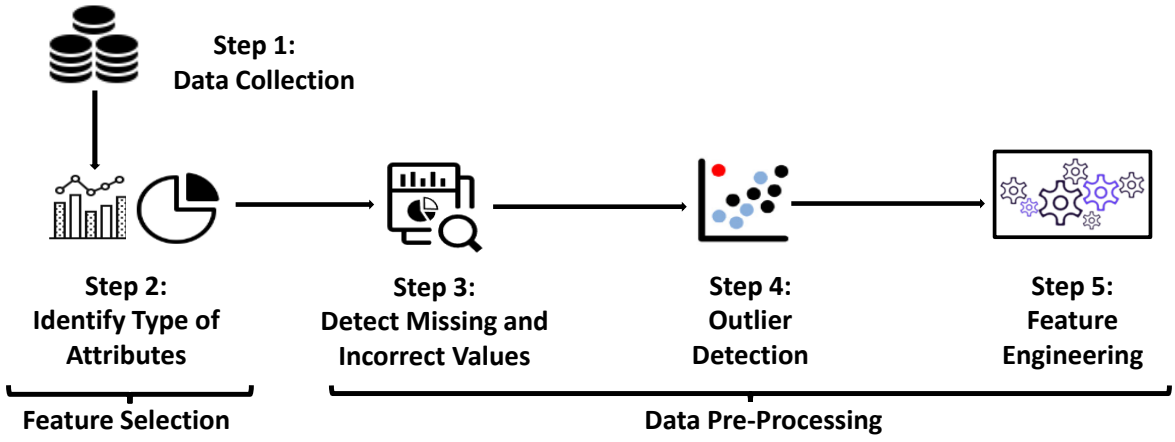

Figure 5.1: The data collection and data processing workflow [8]

certain vocations and regions. Due to the ongoing data collection process and lack of availability of data for all alternative fuel vehicles, the initial study involves investigating the machine learning models to model the maintenance and repair costs in delivery trucks using diesel and natural gas fuels.

5.2 Methodology

The maintenance needs for a vehicle vary based on the vehicle type, fuel type, engine, and many other factors. The vehicles undergo different types of maintenance such as periodic – periodic scheduled maintenance, preventative – check for potential failures and eliminate them, and corrective – after the failure occurred. The cost associated with these types of maintenance varies drastically based on their type, the part of maintenance and the age of the vehicle, and other factors such as region, weather, and the duty cycle of the vehicle. The data collection and data processing workflow are shown in Figure 5.1 and described in the next sections.

5.2.1 Data Collection

The key challenge in studying the influence of vocation, fuel type, and other parameters on maintenance cost is the lack of data. For this study, WVU CAFEE leverages the data collected through a partnership from an ongoing study allowing additional information related to maintenance cost (MC) and fleet-specific maintenance practices of diesel-powered and alternative-fueled

vehicles. The team collected data related to the maintenance of diesel and alternative fuel vehicles from many vendors. The maintenance costs related to three types of maintenance namely preventative, corrective, and periodic are recorded. Each of the maintenance categories is subdivided into major vehicle components such as engine and transmission, tire and brake, fuel system, and exhaust and emissions control system. The current data collected contains 19,641 instances for different vocational trucks. However due to the lack of availability of data for the school bus, transit bus, and refuse trucks using different alternative fuels and the ongoing data collection process, this chapter focuses on modeling data from delivery trucks using diesel and natural gas fuels only. Data has been collected from multiple vehicles where each vehicle underwent several periodic, preventative, and corrective maintenance. The maintenance records include cascaded maintenance where preventative or corrective maintenance is identified during the periodic maintenance or multiple fixings/replacements are performed during a single maintenance.

5.2.2 Feature Selection and Feature Engineering

The data collected has the features shown in Table 5.1. The features such as fuel type, part of truck, engine, region, comments, and season are categorical data types and cannot be fed directly to the prediction models. Hence, these features are converted to numerical data type using one-hot encoding which converts the nominal feature with n different values to n binary attributes [155]. The categorical variable has no ordinal relationship between values hence a new binary value is added for each unique value in the variable. Additional relevant features such as the maintenance number for the vehicle, the time between maintenance (TBM), miles per day, vehicle age, road congestion index (RCI), and cost per mile can be calculated from the original features using a statistical process included in the data. Maintenance number and TBM can give information related to previous maintenance calculated based on the maintenance dates of consecutive maintenance. Miles per day refers to the strength of activity performed by the vehicle before the maintenance has occurred ad is calculated by dividing the difference in mileage by the number of days in between consecutive maintenance. Vehicle Age is calculated as the number of days since the vehicle's first maintenance took place till the date of maintenance. The RCI has been obtained from the bureau of transportation statistics [156]. The cost-per-mile is calculated as the total maintenance cost divided by the number of miles accumulated till the current maintenance. This increased the number of

Table 5.1: Features in the raw data collected

Feature	**Description**	**Feature**	**Description**
VIN	Vehicle Identification Number	Engine	Engine model used in the vehicle
Unit Number	Unique Vehicle Number	Engine Year	Year of Engine Manufactured
Date	Date Maintenance was performed	Make	Make of the Engine
Fuel Type	Type of fuel used in vehicle	Region	Region of Operation of Truck
Mileage	Mileage at which maintenance has occurred	Repair Shop Name	Name of the shop where maintenance has taken place
Part of Truck	Part of the Truck	Comments	Type of Maintenance
Total Cost	Cost of the maintenance during that visit	Season	The season during vehicle maintenance
Owner of truck	Owner of the operating truck	Vocation	Activity performed by the vehicle

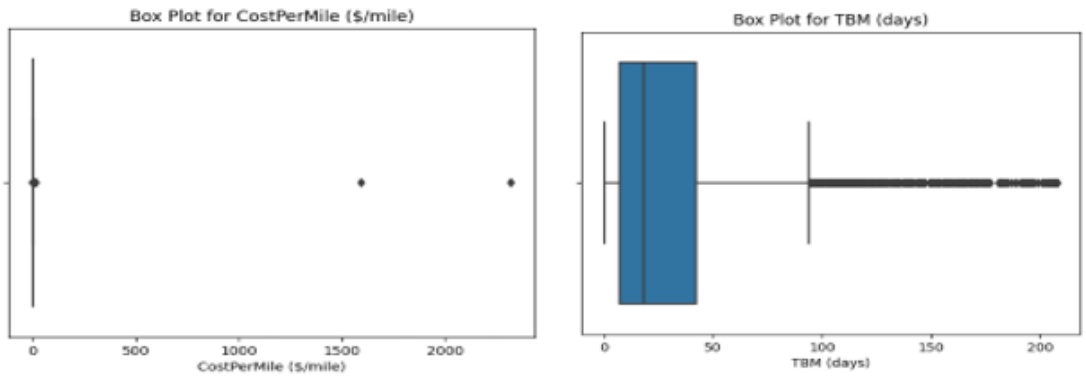

Figure 5.2: Outlier Plots

input features in the dataset from 9 to 13 with cost-per-mile as the target variable. However, the effect of these derived features on the maintenance cost is unknown.

5.2.3 Data Pre-Processing

The data contains 19,641 instances for all vocational trucks. However, there has been a lot of duplicate data with varying total costs, hence the data has been consolidated by adding the instances of the same vehicle having the same data in all features except the total cost. This reduced the number of instances to 8,378 of which 1,412 instances are diesel delivery trucks, and 3,690 instances are natural gas delivery trucks. Since there is more data available for delivery trucks, the initial study concentrated on modeling the data related to delivery trucks and later extend it to include different vehicle types.

Firstly, the data with missing values and extreme values are identified which may be caused by the error in the manual entry or abnormal odometer reading. Records containing negative values or missing values are removed for reliability. Outlier analysis is performed using the Z-Score method to identify instances that deviate from the distribution, but the outliers may also provide some key information. Based on the box plots for IQR, a careful inspection of data points outside the whiskers is performed to determine whether the data points should be considered an outlier. For example, the case when maintenance incurs a high cost due to a large part replacement is labeled as an outlier (shown in Figure 5.2), however, it carries important information related to corrective maintenance. Such data points are included for modeling after careful analysis.

After all the pre-processing is applied, there are 4,910 instances in the dataset that are robust for analysis. To understand the correlation of features, a correlation plot shown in Figure 5.3 is used. The feature association is about the correlation or the level of dependence of one feature on another. The correlation score tells association strength which is how accurately a feature can be determined based on the other. A correlation score of 0 means the features are independent. These instances are indicated with lighter dots on the feature association matrix in Figure 5.3. A correlation score of 1 indicated perfect correlation and is represented by darker dots on the matrix. The features that are correlated are clustered into groups seen as blue, green, and orange on the matrix. The darkness of the dots in these clusters further indicates how strongly the features are correlated. Ideally, multicollinearity should be avoided, as the high degree of correlation can cause a problem in fitting a model.

In addition to the removal of outliers and analysis of feature correlation, the data are also randomized to make machine learning models capture patterns in data. Without randomization, machine learning methods applied to data may predict well short-term maintenance well, but not the one that has higher vehicle age with larger mileage. Therefore, the randomization of data is important in the machine-learning process. With the numerical data, different features have different ranges of values as the features are not unified. To protect data integrity, normalization is performed. The dataset is then divided into training and testing sets.

5.2.4 Machine Learning Analysis

Deep learning (DL), a class of machine learning algorithms, has demonstrated an exclusive performance in application to data containing images and videos. However, on tabular data, analyzing the performance of DL methods presents a challenge. Therefore, machine learning ensemble models remain the promising tool as they do not require that objective functions be differentiable, an inductive bias be introduced, and a large sample size for training [24]. In addition, deep learning neural networks on tabular data require a lot of data pre-processing and exploratory data analysis as data containing irrelevant features that make the neural network performance worse especially with irregular patterns in the target. Most importantly, machine learning algorithms are easy to train and have low tuning costs. However, the data collected are imbalanced as there is a much larger number of instances of diesel vehicles than other alternative fuels since the usage of alterna-

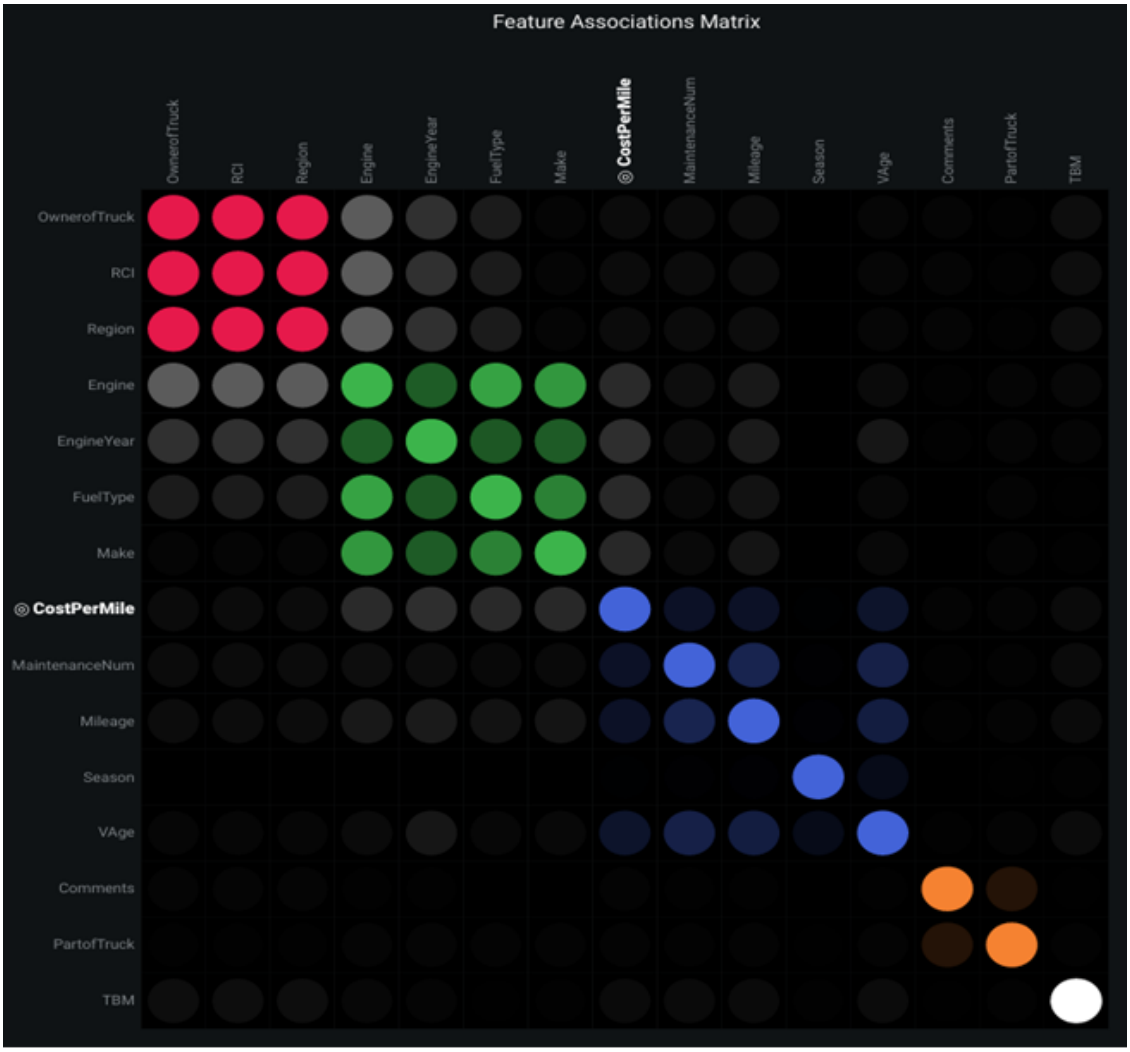

Figure 5.3: Feature Association Matrix

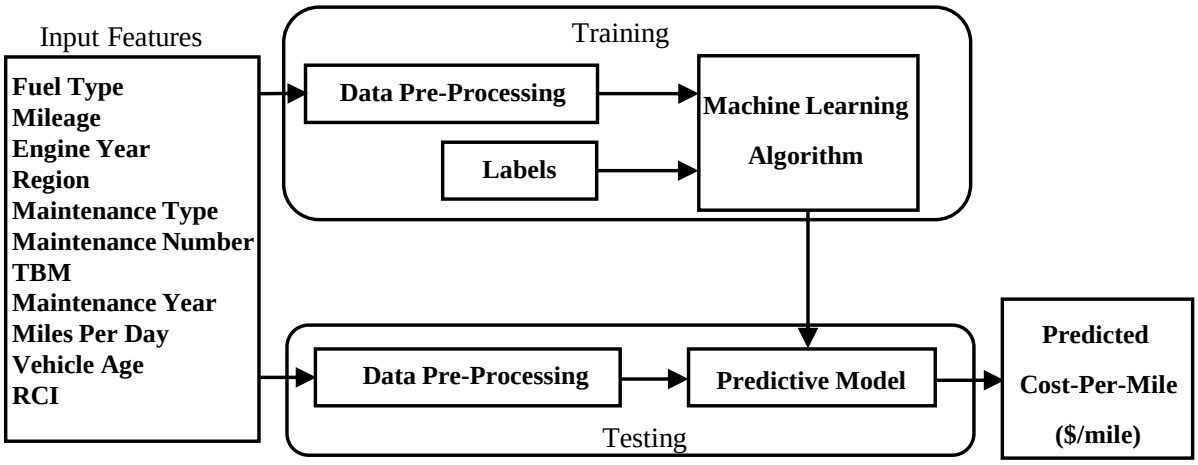

Figure 5.4: Machine Learning Workflow

tive fuels is very low in the heavy-duty vehicle sector [157] resulting in a lack of data. To prevent bias in the performance of the machine learning models in estimating the maintenance cost, the initial study only considered the maintenance data from delivery trucks using diesel and natural gas.

Various machine learning algorithms for predicting continuous variables are investigated on the diesel truck maintenance data. The general process of machine learning prediction models is shown in Figure 5.4. The selected features from the data collected along with the derived features are passed to the machine learning model for training. The model is then used to make predictions on unseen data.

In this study, several machine learning algorithms [158]: multiple linear regression [125], support vector machine (SVM) [140], k-nearest neighbor (k-NN) [159], random forest (RF) [119], extreme gradient boosting (XGB) [160], and a neural network (NN) are evaluated on the pre-processed data. Each machine learning model has its hyper-parameters that need to be determined to achieve the best-performing model. A grid search has been performed for each of the algorithms to identify the hyper-parameters. Since the amount of data is limited, k-fold cross-validation is performed to prevent overfitting and to understand how well the model generalizes for unseen data. The Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2) are widely used in regression tasks to compare the performance of models. The

mathematical expressions for RMSE, MAE, and (R^2) are given below

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|x_i - \hat{x_i}|, \tag{5.1}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x_i})^2}, \tag{5.2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(x_i - \hat{x_i})^2}{\sum_{i=1}^{N}(x_i - \overline{x})^2}, \tag{5.3}$$

where x_i is the actual measured fuel consumption at time i, $\hat{x_i}$ is the predicted value at time i, $\overline{x}$ is the mean of data given by $\sum_{i=1}^{N} x_i/N$, and N is the number of data points.

5.2.5 Multiple Linear Regression

Multiple Linear Regression is the most well-known regression technique where the data is fitted to a straight line to predict output by minimizing a cost function or error. In this study, a multi variable linear equation given by equation 5.4 is used due to multiple input parameters.

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 \tag{5.4}$$

Where, y is the output and x_1, x_2, x_3 are the input variables with $\theta_0, \theta_1, \theta_2, \theta_3$ being parameters to learn.

Support Vector Machine (SVM)

SVM is a state-of-the-art supervised machine learning model used for classification and regression analysis. The kernel function such as polynomial, radial basis function, sigmoid, etc. used in SVM makes it effective in non-linear mapping by mapping inputs to high-dimensional feature spaces. Different kernel functions were tested, and it was found that for the given data, the poly kernel suits the best.

K-Nearest Neighbor (k-NN)

K-NN is another simple supervised instance-based machine learning algorithm for performing both classification and regression tasks. It is referred to as lazy learning as the algorithm does not create a model but instead uses all the data instances to calculate the similarity measure. However,

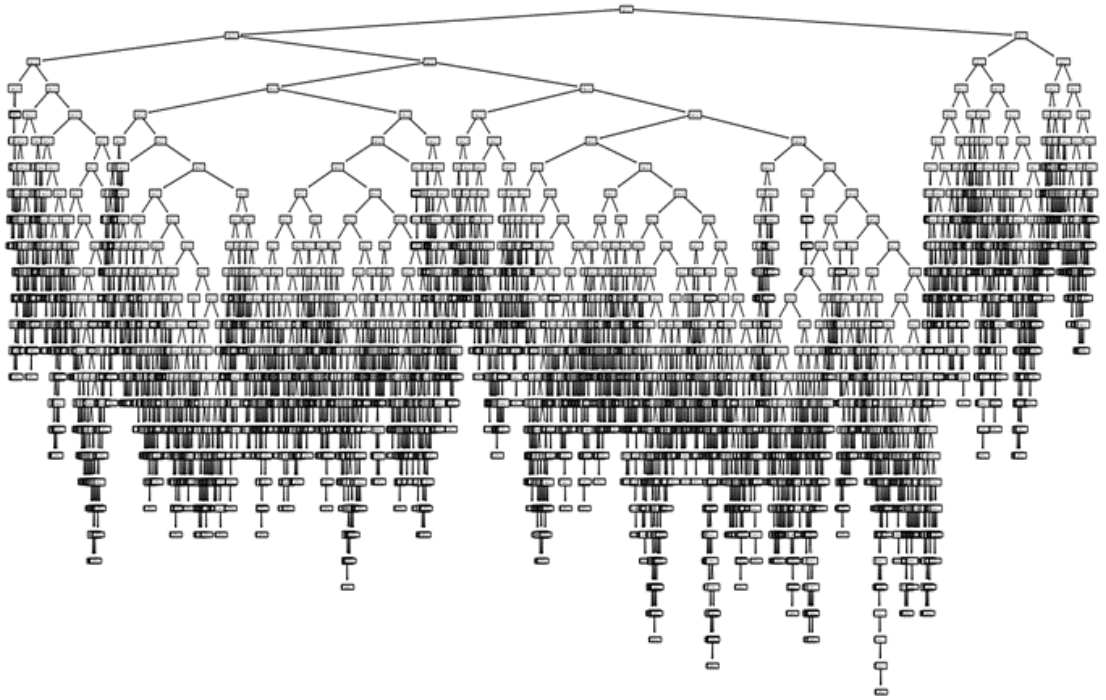

Figure 5.5: Random Forest Model developed for Maintenance Cost Data

its major drawback is that the number of neighbors and the performance is highly dependent on the quality of the data. In addition, this algorithm can be slow when applied to large datasets as it takes all instances into memory. With the hyper-parameter search number of neighbors was determined to be 5 with uniform weight and Euclidean distance.

Random Forest (RF)

RF is an ensemble machine-learning technique that uses decision trees to build a model. The model fits several decision trees built in parallel from subsets of data. The final decision is based on the average for regression tasks. Since it internally performs cross-validation, over-fitting is minimized. The key hyper-parameter in the performance of RF is the number of trees which is set to be 25 in the model developed. The developed RF model is shown in Figure 5.5.

Extreme Gradient Boosting (XGB)

Like RF, XGB is an ensemble learning algorithm that considers more detailed approximations to build the final model based on a series of decision trees. The over-fitting is handled by minimizing

Figure 5.6: XGBoost Model developed for Maintenance Cost Data

the loss using gradients. XGB performs well with larger datasets, and its training is fast due to parallel processing. The developed model (shown in Figure 5.6) has a learning rate and the number of trees as the key parameters that are set to 0.1 and 100, respectively.

Neural Network (NN)

NNs are referred to as a system imitating the operation of human brain neurons. A NN or a multi-layer perceptron (MLP) is a base architecture of deep learning typically containing an input layer, hidden layers, and an output layer with multiple neurons called a perceptron. The network learns by updating the weights of neurons through the backpropagation technique. NNs may become computationally expensive and require lots of hyperparameter tuning such as the number of neurons, number of layers, learning rate, batch size, etc. A 6-layer fully connected dense network with 4 hidden layers containing 15 neurons in each layer is developed. The input layer contains 10 inputs, and the output layer contains single neurons. All the neurons are activated with an activation function called ReLU (Rectified Linear Unit) and an Adam optimizer with a learning rate of 0.001 is used. The model is trained using a batch size of 64 for 200 iterations and the mean square error loss function. All the hyper-parameters are selected after several tests and grid searches. The developed NN is shown in Figure 5.7.

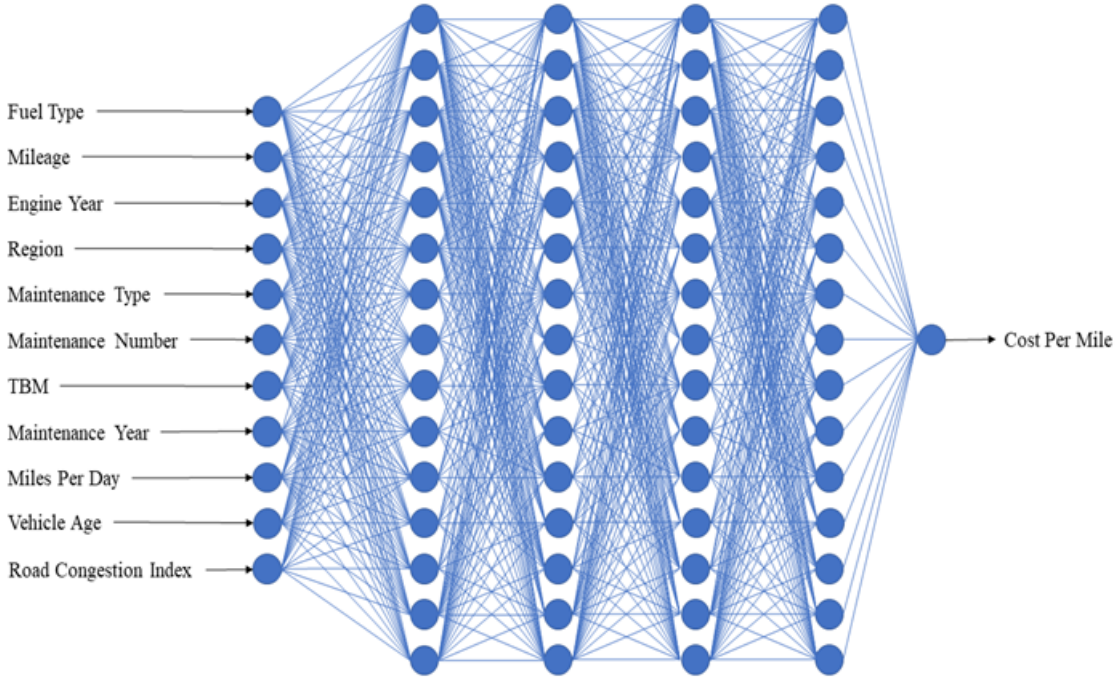

Figure 5.7: Neural Network Model developed for Maintenance Cost Data.

Super Learner Model

The Super-Learner model [59] is developed by stacking multiple base algorithms to improve the performance of the model. In this work, a super-learner model based on the predictions from base models such as random forest (RF) and XGBoost (XGB) and a linear regression meta-model is developed. The framework for super-learner is shown in Figure 5.8. The training of the super-learner model involves various steps. The dataset is initially divided into training and validation datasets. A 5-fold cross-validation is performed using the training dataset to prevent the over-fitting of the model whereas the validation dataset is used for the validation of the model. The folded training data are used to train the base models (random forest and XGBoost) and the validation set is passed to the base model to make the predictions. The predictions on the test fold of the training data are stacked for a new training dataset and the averaged validation dataset outputs are stacked to get a new validation dataset. The new datasets are weighted based on the performance of the base models to add more significance to the best-performing base model. The super-learner (meta model) is then trained with the weighted new training dataset and evaluated using the weighted new validation set. This ensemble technique performs as well as or better than

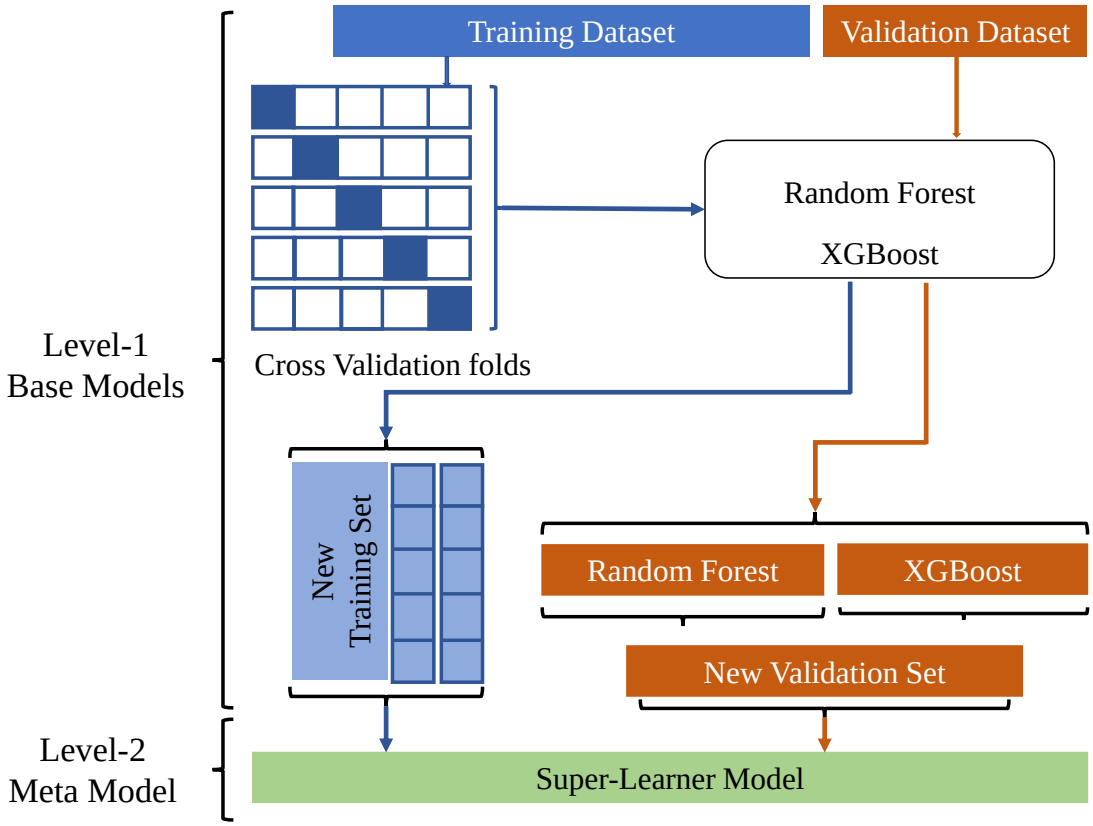

Figure 5.8: Super Learner Model [9]

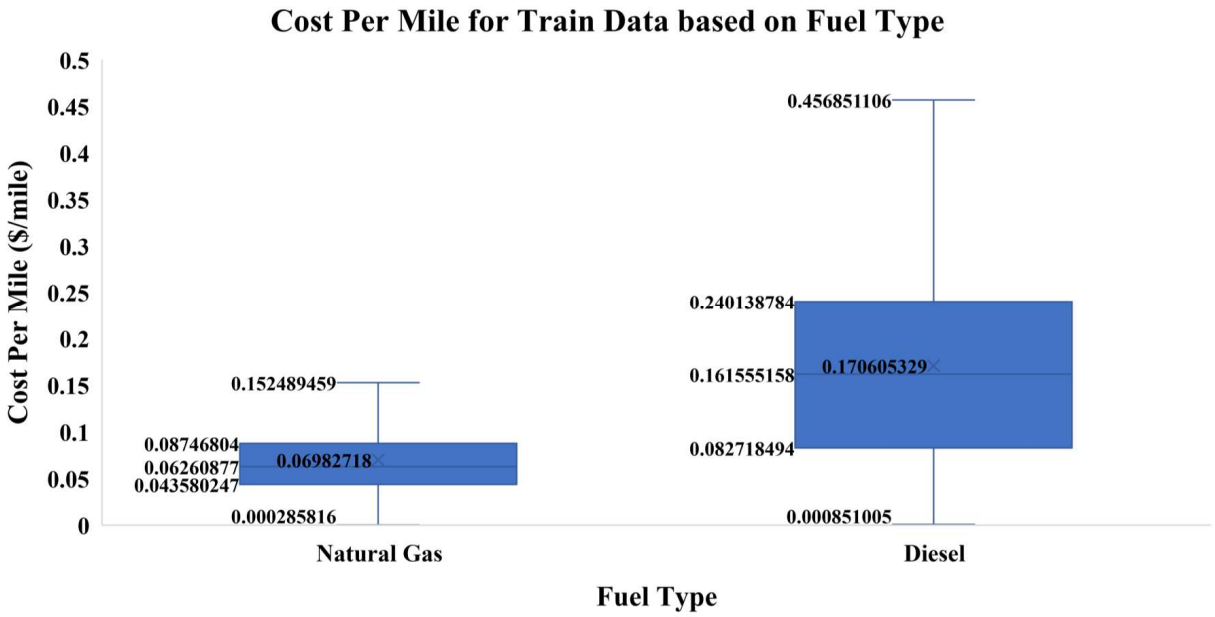

Figure 5.9: Distribution of cost per mile ($/mile) vs fuel type for train data

an individual model. The super-learner model improves the prediction capability of the model and is more robust to overfitting and underfitting.

5.3 Results and Discussion

Delivery trucks are prominently used in transporting goods whose lifespan ranges from 50,000 miles to 150,000 miles and even more. The data used in this study included 18 delivery trucks fueled by diesel and 12 delivery trucks fueled by natural gas. The diesel trucks ranged in mileage from 2896 to 895592 miles while the natural gas trucks ranged in mileage from 756 to 563249 miles. The model year of diesel trucks was in the range of 2010 to 2022 while that of natural gas was in the years 2014 to 2017. The data includes various types of maintenance such as periodic, corrective, and preventative for major parts such as chassis, engine and transmission, tire and brake, exhaust and emission, and fuel system. Natural gas vehicles have a lower cost per mile ranging from 0.000285 - 0.152489 $/mile with few outlier points whereas diesel fuel vehicles have a cost per mile ranging from 0.000851 - 0.456851 $/mile. The higher value is the indication of corrective maintenance for replacing the failed larger or complex parts such as transmission, exhaust system, etc. that have incurred a very large amount. Since these data points are important for calculations, they are

Table 5.2: Comparison of cross-validation scores for all ML models using original features

Model	**MAE ($/mile)**	**RMSE ($/mile)**	R^2 **Train**	R^2 **Validation**
XGB	0.0106 ± 0.0005	0.0172 ± 0.0022	0.9754 ± 0.0007	0.9400 ± 0.0169
RF	0.0127 ± 0.0003	0.0199 ± 0.0011	0.9820 ± 0.0005	0.9306 ± 0.0115
k-NN	0.0150 ± 0.0005	0.0252 ± 0.0009	0.9201 ± 0.0017	0.8735 ± 0.0102
NN	0.0217 ± 0.0031	0.0329 ± 0.0037	0.8081 ± 0.0564	0.7821 ± 0.0462
MLR	0.0320 ± 0.0017	0.0500 ± 0.0025	0.6896 ± 0.0194	0.6735 ± 0.0289
SVM	0.0580 ± 0.0021	0.0623 ± 0.0016	0.9704 ± 0.0007	0.9301 ± 0.0169

included in the analysis. Diesel vehicles include an after-treatment system that consists of multiple catalytic systems that require increased maintenance to reduce vehicle downtime. Whereas natural gas vehicles compared to diesel have only a single catalytic system in the form of a three-way catalyst (TWC) offering lower maintenance costs, especially the corrective maintenance cost as shown in Figure 5.9. The results for the models on the original data without the calculated features are shown in Table 5.2. All the results are obtained by 5-fold cross-validation. The cross-validation scores represent the average ± standard deviation of the results from 6 models trained by randomly splitting the dataset 5 times. The MAE and RMSE are used to evaluate the performance of the regression model while R^2 explains how well the variability in data is explained by the model. The lower values of MAE and RMSE quantify how well the model can predict. From Table 5.2, XGB has the smallest values of MAE and RMSE, and the highest values of R^2 for both the training and testing sets compared to other models. XGB is followed by RF. The third lowest error is observed for k-NN but as the data increases, k-NN requires more memory and time to process data. Unlike other models, k-NN depends on the neighboring samples using feature similarity to predict the values of unseen data rather than creating a model. Even though NN has a lower coefficient of determination, the error values are lower compared to SVM. Although SVM has a higher coefficient of determination, the error rate is high, hence SVM might not be generalizing well for unseen data. Considering the training time, memory requirements and the need for understanding data, XGBoost and RF are the best-performing models with the NN following them. If we had more data to train and test NN, it may perform better than its current version.

The second set of training involved the newly derived features in the input features as described in section 5.2.2. Table 5.3 shows the results for all the machine learning models with the new

Table 5.3: Comparison of cross-validation scores for all ML models using combined features

Model	**MAE ($/mile)**	**RMSE ($/mile)**	R^2 **Train**	R^2 **Validation**
XGB	0.0092 ± 0.0003	0.0151 ± 0.0018	0.9920 ± 0.0003	0.9538 ± 0.00123
RF	0.0111 ± 0.0004	0.0180 ± 0.0013	0.9955 ± 0.0003	0.9353 ± 0.0091
KNN	0.0136 ± 0.0003	0.0235 ± 0.0009	0.9299 ± 0.0015	0.8901 ± 0.0091
NN	0.0202 ± 0.0012	0.0292 ± 0.0011	0.8451 ± 0.0084	0.8302 ± 0.0130
MLR	0.0294 ± 0.0007	0.0428 ± 0.0006	0.6941 ± 0.0194	0.6903 ± 0.0018
SVM	0.0611 ± 0.0013	0.0650 ± 0.0011	0.9920 ± 0.0003	0.9538 ± 0.0124

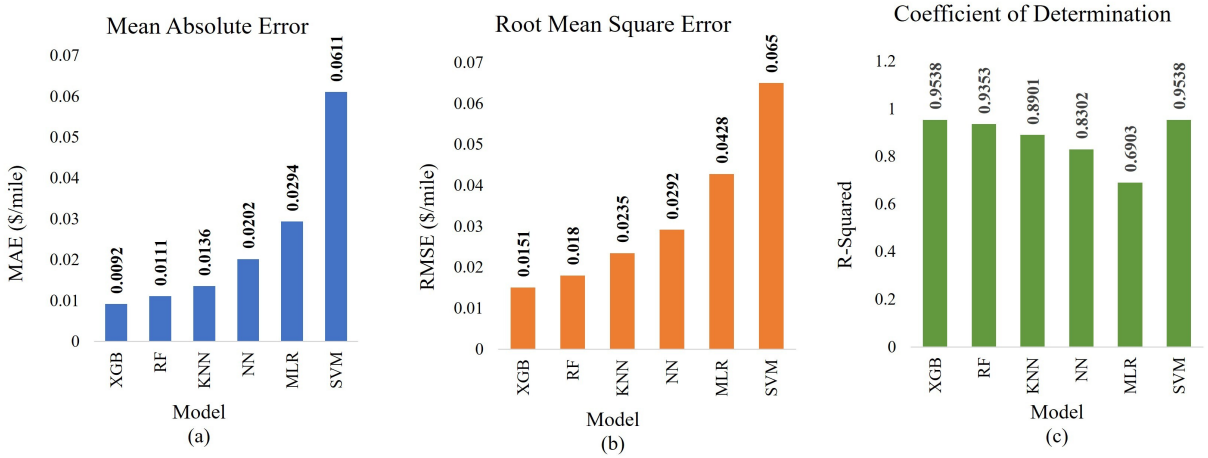

Figure 5.10: Comparison of Cross Validation Performance for ML models. Mean Absolute Error (Left Panel), Root Mean Square Error (Center Panel), Coefficient of Determination (Right Panel) using machine learning models.

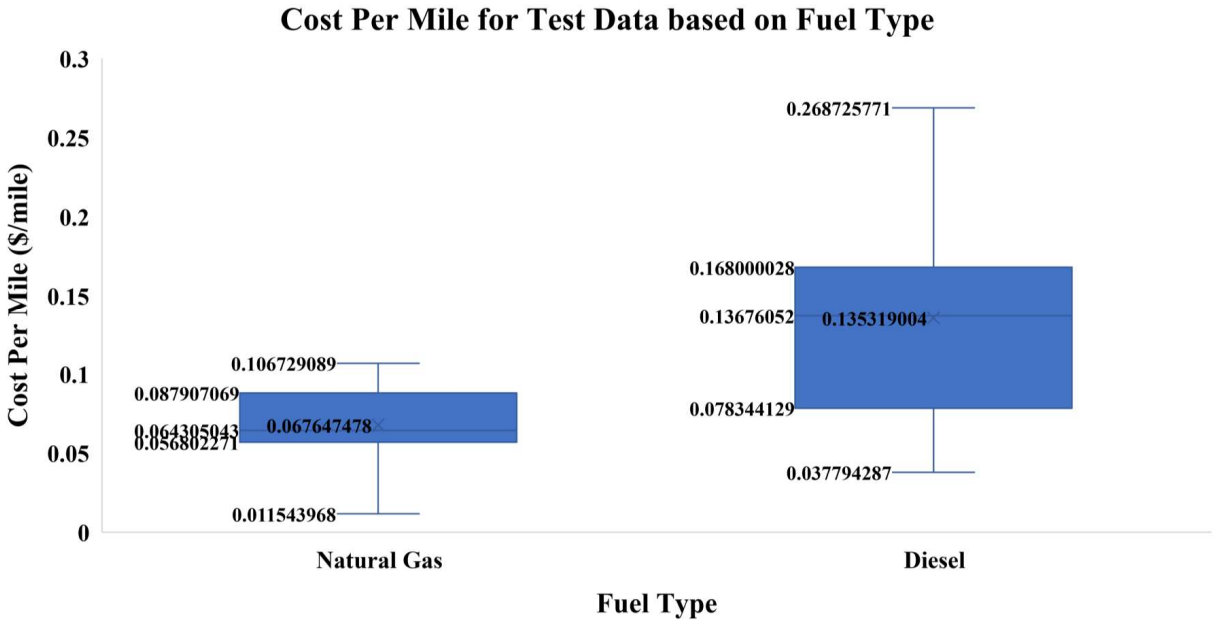

Figure 5.11: Distribution of cost per mile ($/mile) vs fuel type for test data

combined feature input. In this case, the error for most of the models is lower than in the previous case and the coefficient of determination has increased.

The cross-validation performance metrics presented in Table 5.3 is plotted in Figure 5.10. Based on the cross-validation results, the tree-based models XGB, RF and the SVM have higher R^2 but the MAE and RMSE using SVM is high whereas for tree-base models the error is less.

However, to complete the performance evaluation, the models need to be evaluated using the testing dataset. Two vehicles using diesel and natural gas fuels operated with 5 and 7 years of maintenance records respectively are used as test data. The average cost-per-mile distribution of the testing data for diesel and natural gas is shown in Figure 5.11. Like the distribution of the training data, the original testing data show that natural gas has a lower average cost per mile than diesel fuel trucks. The natural gas vehicles have a lower cost per mile ranging from 0.011543 - 0.106729 $/mile with few outliers whereas the diesel fuel vehicles have an average cost per mile ranging from 0.037794 - 0.268725 $/mile.

The model is evaluated using evaluation metrics for all the models on test data as shown in Figure 5.12. Based on the cross-validation and the model evaluation results, the tree-based ensemble model performs better on the tabular data compared to neural networks. Though neural

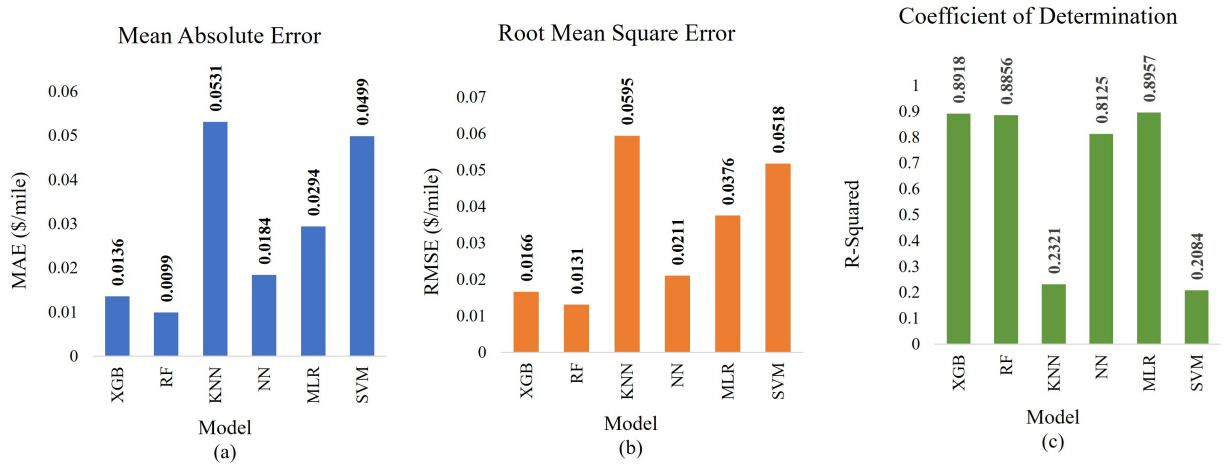

Figure 5.12: Comparison of ML model performance on test data. Mean Absolute Error (Left Panel), Root Mean Square Error (Center Panel), and Coefficient of Determination (Right Panel) using machine learning models.

networks have gained success on data containing images, videos, text, audio, etc. their performance on tabular data is still unclear and machine learning ensemble models remain the promising tool as they don't need differentiable functions [24]. Neural networks on tabular data require a lot of data pre-processing and exploratory data analysis as data containing irrelevant features makes the neural network performance worse especially with irregular patterns in the target. Most importantly, Machine learning algorithms are easy to train and have low tuning costs.

Based on the performance evaluation metrics for each model, it is observed that SVM and KNN have very low R^2 values for test data followed by neural networks unlike the train data indicating that these models do not generalize well for unseen data. This might be due to an imbalance in the train data. Moreover, neural networks require more training data and training time and are computationally expensive. Though the multiple linear regression has an R^2 value of 0.9457, the model assumes a linear relationship between the features and does not explain well if the relationship is not linear. The tree-based ensemble models random forest and XGBoost seem to generalize well for unseen test data.

The scatter plot for actual and predicted cost per mile for the test dataset using the machine learning models trained is shown in Figures 5.13 and 5.14. The red diagonal line is where the actual value and predicted value are equal. The blue data points along with the blue line represent the data with the regression fit. The closer the red line to the blue line is in terms of the distance and

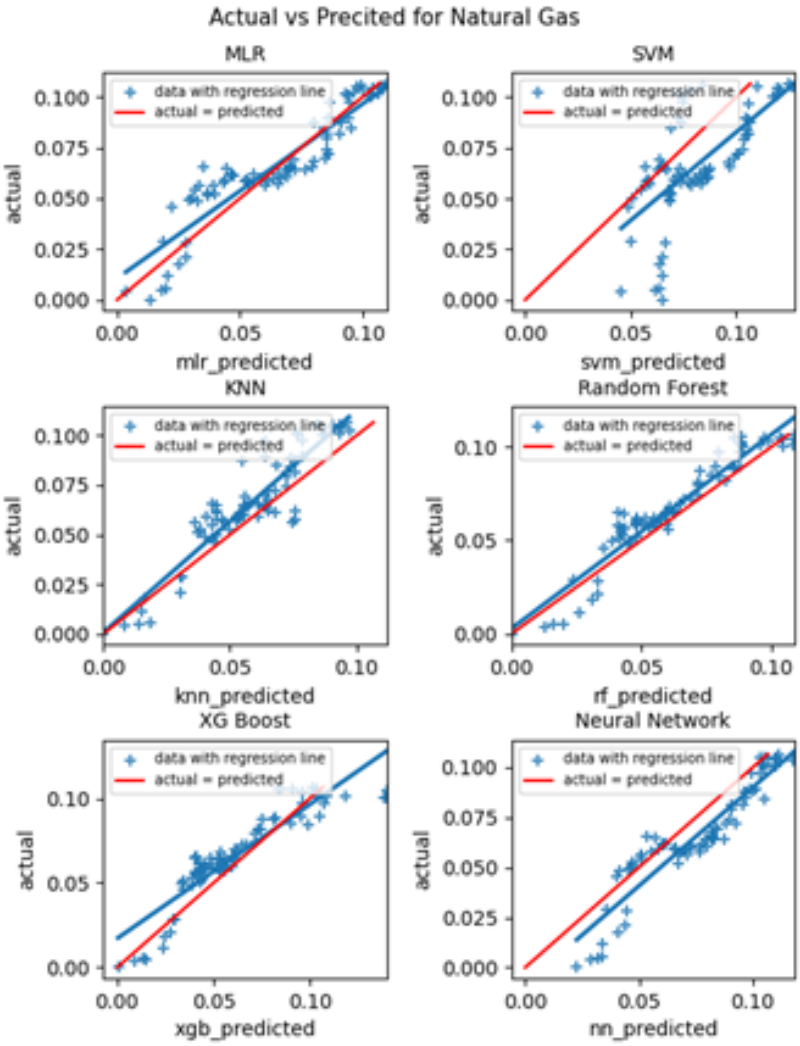

Figure 5.13: Scatter plots for actual and predicted cost per mile based on fuel type Natural Gas for test data using different machine learning models

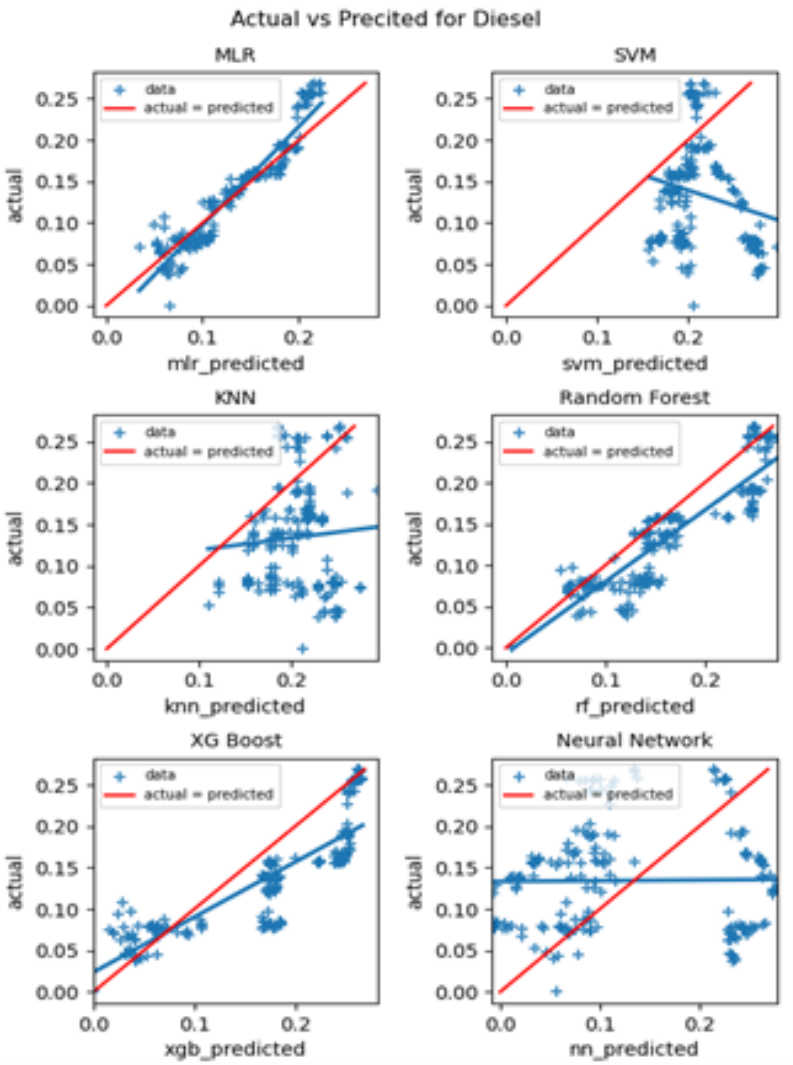

Figure 5.14: Scatter plots for actual and predicted cost per mile based on fuel type Diesel for test data using different machine learning models

Table 5.4: Cross-validation results for Super-Learner Model

Model	**CV MAE ($/mile)**	**CV RMSE ($/mile)**	R^2 **Train**	R^2 **Validation**
XGB	0.0092	0.0151	0.9501	0.9792
RF	0.0110	0.0180	0.9421	0.9755
Super Learner	0.0068	0.0086	0.9728	0.9880

slope, the better the model performs. SVM and k-NN models' regression lines substantially deviate from the true performance lines making the models poor options. Each of the RF and XG boost models has a regression line slightly deviating to the left from the actual line indicating that these models perform well but they are predicting slightly lower than the actual values should be. The neural network has a regression line to the right of the actual line indicating the over-predicted values. The same trend is observed among models for the diesel data, however, the fitted lines have much larger deviations from the actual lines compared to the case of Gas fuel. This trend might be because the training dataset has less data related to diesel vehicles than natural gas.

The random split in data, however, tends to change the feature importance due to an imbalance in data for each machine learning model as few maintenances occur rarely and each vehicle is maintained differently based on usage, ownership, driving behavior, etc. As each machine learning model works differently and tends to have different feature importance as shown in Figure 5.15 the performance of machine learning models is affected.

Hence, a super learner ensemble model, shown in Figure 5.8 is developed. The super learner model has two levels where the prediction from individual base models in level 1 is used for training the level 2 meta-model. Using random forest, and XG boost as base models and fusing the results using a linear regression model improves the performance leading to cost-effective maintenance in both diesel and natural gas vehicles. The super-learner ensemble model works by training the meta-learning algorithm by creating an optimal weighted average of the level one base learners per k-fold cross-validation to approximate complex relationships. Super-learner eases the issue of selecting the right learners for learning a function. The performance metrics for the super learner and the base learners on the training dataset are shown in Table 5.4. The ensemble model achieved a mean absolute error as low as 0.0068, a root mean square error of 0.0086 with R^2 equal to 99.79%

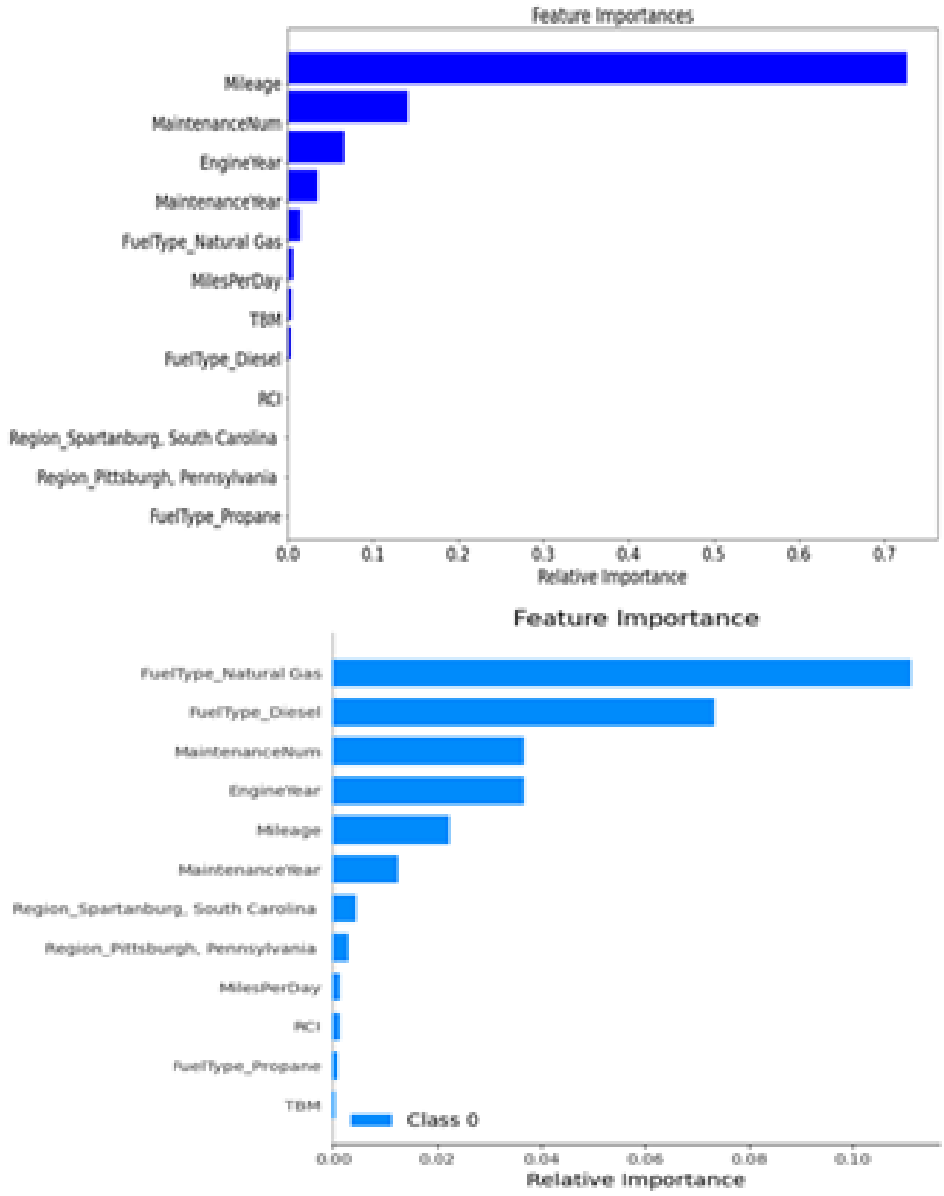

Figure 5.15: Feature Importance using Random Forest (Top Panel) and Neural Network (Bottom Panel)

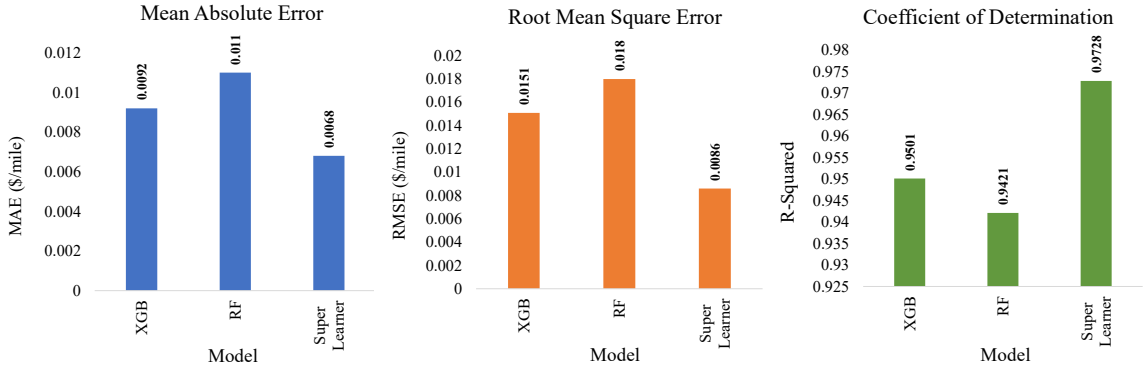

Figure 5.16: Cross Validation Performance using Super-Learner Model. Mean Absolute Error (Left Panel), Root Mean Square Error (Center Panel), Coefficient of Determination (Right Panel)

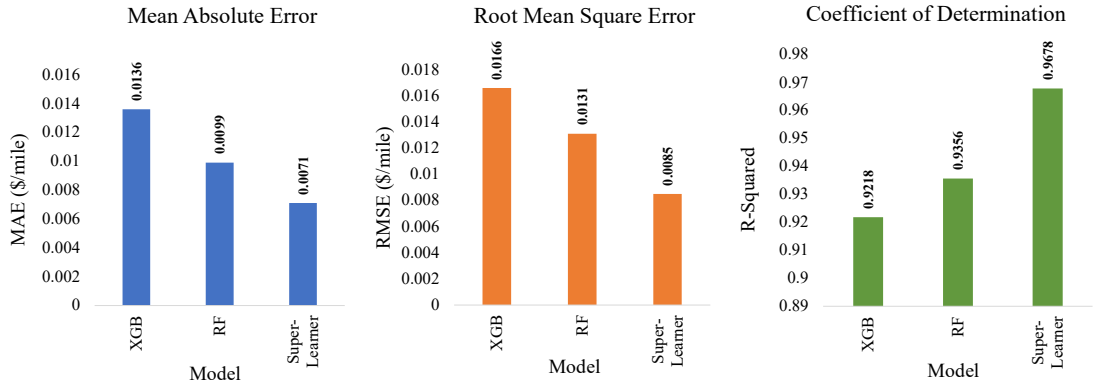

Figure 5.17: Model Evaluation on Test Data using Super-Learner Model. Mean Absolute Error (Left Panel), Root Mean Square Error (Center Panel), Coefficient of Determination (Right Panel)

for training data and 95.46% for the validation dataset which is better than the results seen in Table 5.3.

The comparison of cross-validation performance metrics for the super-learner model and the base models used in super-learner model presented in Table 5.4 is plotted in Figure 5.16. Based on the plots, the performance of super-learner model is higher compared to the base models, random forest and XGBoost. The super-learner performance is then validated by testing the model on test data. The performance of super-learner model along with the base models is shown in Figure 5.17.

The performance of super-learner model on the entire testing dataset containing data for diesel and natural gas delivery truck seems good with R^2 of 96.78%, MAE of 0.0071 ($/mile) and RMSE of 0.0085 ($/mile). To see the generalization for each individual fuel type, the scatter plot for actual

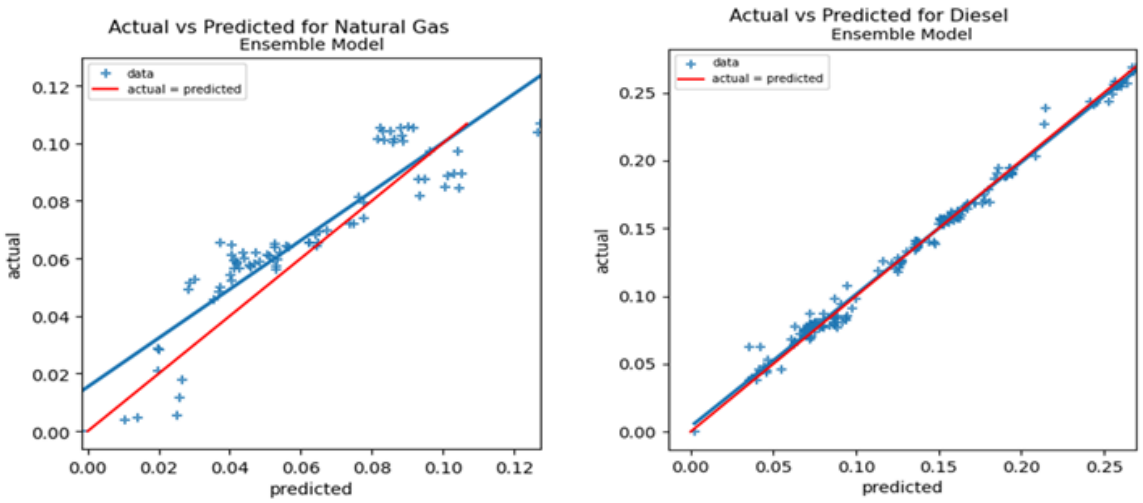

Figure 5.18: Scatter plot for actual and predicted cost per mile based on fuel type Natural Gas (Left Panel) and Diesel (Right Panel) for test data using super-learner ensemble model

and predicted cost-per-mile is plotted in Figure 5.18. Unlike previous models, the super learner model demonstrates better or similar prediction capability for both natural gas and diesel data as the actual and predicted lines are almost aligned.

The average cost per mile over the age of maintenance is shown in Figure 5.19. The results for natural gas vehicle data using the super-learner model seem to be close to random forest model performance whereas the results for diesel vehicle data are close to the actual values unlike before. Because of the prominent usage of diesel vehicles, the maintenance duration for diesel vehicles is longer than other alternative fuel vehicles. However, from the data and the results it is evident that the average maintenance cost for diesel vehicles is higher than for natural gas vehicles for a given lifetime.

The predicted average cost per mile over 5 years for natural gas and diesel vehicles are compared in Figure 5.20. As the original data indicate, the predicted average cost per mile for natural gas is lower than for diesel vehicles. The resulting model serves as a tool for fleet companies to estimate maintenance costs. This helps the companies understand how a new vehicle purchased for a particular task is going to affect the cost spent on its maintenance and repairs. However, when data related to different vocational vehicles are considered, these models fail to generalize well, hence more robust models are required to capture the functions in data.

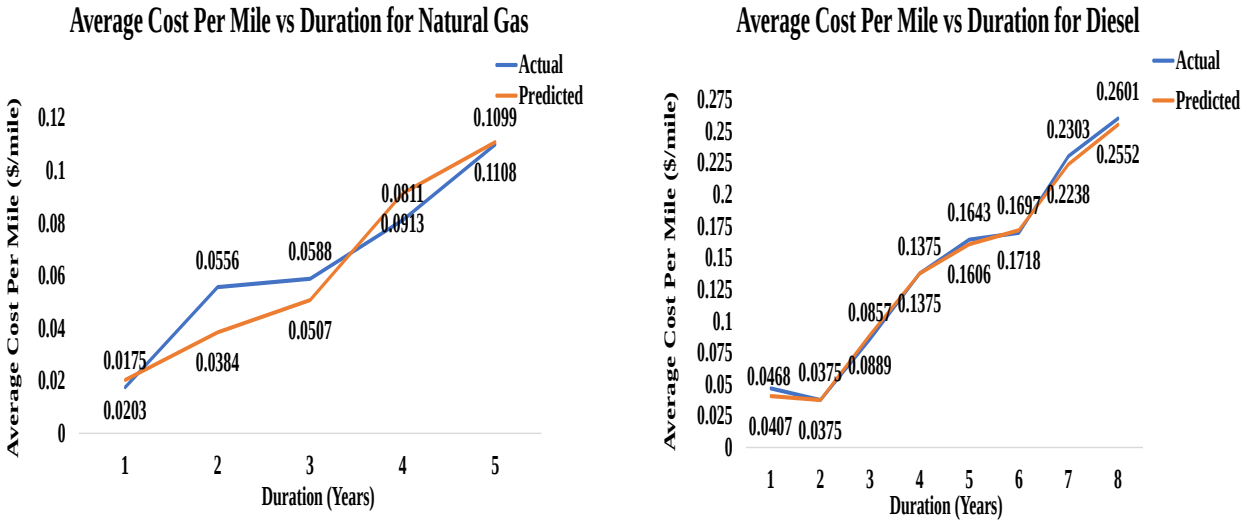

Figure 5.19: Average Cost per mile over Years for Natural Gas (Left Panel) and Diesel (Right Panel) Test Vehicles using Super-Learner Ensemble Model

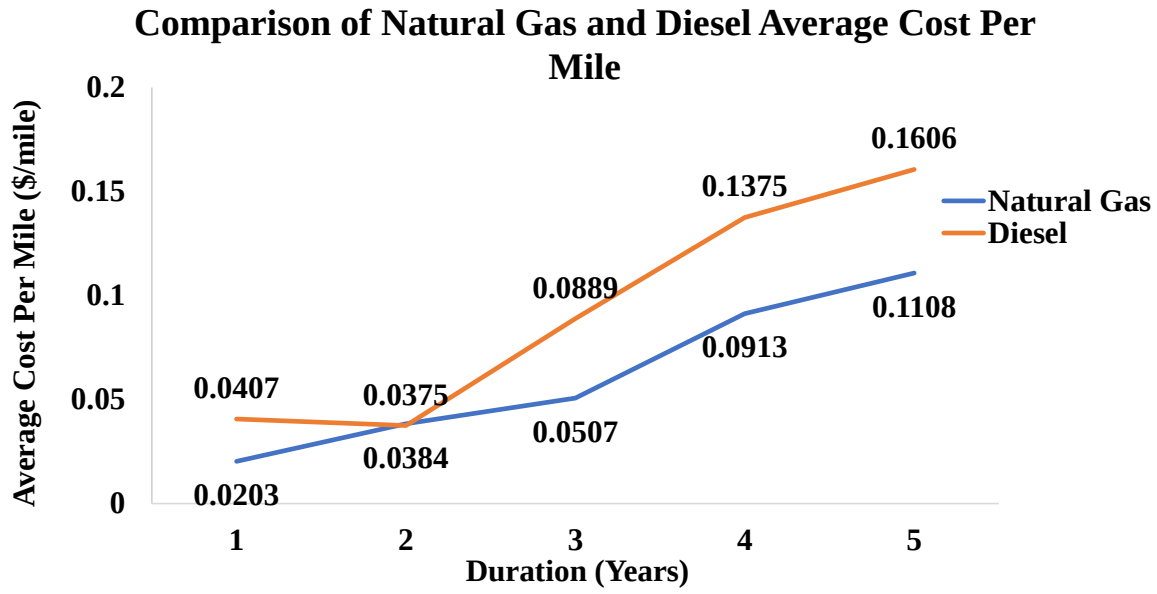

Figure 5.20: Comparison of Predicted Average Cost Per Mile over Vehicle Age for Natural Gas and Diesel Delivery Trucks using the Super-Learner Model

5.4 Conclusion

Maintenance Cost is one of the important considerations for fleet companies. Understanding how the maintenance cost is changing over the years given the vehicle operation, fuel type, region of operation, etc., would enable the fleet companies to make data-driven decisions on the procurement of vehicles that reduce their overall total cost of ownership. This chapter presented an extensive literature survey on the topic of data-driven fleet maintenance solutions that have been published in recent years, for the commercial vehicle industry. One of the major gaps found in the literature is the availability of real-world maintenance datasets from a diverse set of vocational applications, which has limited the validation of algorithms developed using ML techniques. The present study aimed to address this gap by using real-world datasets collected from a variety of fleet companies.

In this work, an investigation study based on real-world maintenance data using machine learning algorithms to predict the average cost per mile is shown. Different machine learning algorithms such as multiple linear regression, support vector machines, k-nearest neighbors, neural networks, random forest, XGBoost, and super-learner models were applied to the well pre-processed data. A 5-fold cross-validation technique is performed to understand the generalization of the model with comprehensive results. To evaluate the performance of models several regression metrics such as mean absolute error (MAE), root means square error (RMSE), and coefficient of determination (R^2) were used. The Super-learner model has demonstrated promising results, and the comparison of predicted values for delivery trucks using diesel and natural gas fuel types has been shown. The super-learner works by training the meta-learning algorithm by creating an optimal weighted average of the level one base learner per k-fold cross-validation to approximate complex relationships. Super-learner eases the issue of selecting the right learners for learning a function. The super-learner model achieved errors as low as 0.0068 $/mile for mean absolute error (MAE) and 0.0086 $/mile for root mean square error (RMSE) with a coefficient of determination (R^2) equal to 99.8%. This study is important as it fills the gap by considering real-world maintenance and vehicle activity data and compares natural gas alternative fuel with diesel fuel in delivery trucks. The results demonstrated lower maintenance costs for natural gas delivery trucks, indicating that opting for alternative fuel vehicles offers a low cost of ownership along with reduced emissions. The outcome of the solution developed in this study could potentially pave the way for creating a cloud-

based application, that can ingest the data from fleet companies, run the deployed machine learning model in the cloud, and enable the companies in making data-driven decisions for procurement of newer fleet that will have a positive impact on meeting their TCO targets and sustainability goals, by adopting alternative fuel vehicles.

Chapter 6

Mixed Effects Model for Estimating Maintenance Costs in Heavy-Duty Vehicles using Alternative and Diesel Fuels

Transportation plays a vital role in socioeconomic interactions, especially freight transportation. The poor planning of regular maintenance, driving conditions such as road, driver behavior, increased fuel costs, etc. impact the reliability of vehicles resulting in frequent breakdowns. These frequent breakdowns increase the downtime of the vehicle as well as maintenance costs.

Along with fuel and labor costs, maintenance costs are significant for companies. In the previous chapter, the maintenance data for delivery trucks using diesel and natural gas fuels were modeled using machine learning models and a super-learner model for estimating the maintenance costs. However, these models do not perform well when data of different vehicles such as school buses, good movement, transit buses, refuse trucks, and vocational trucks operating on different fuel types such as diesel, natural gas, and propane are mixed together. This is due to the clustered nature of data collected from vehicles that perform different activities. Each heavy-duty vehicle type has a different duty cycle based on the activity it performs. For example, a refuse truck mostly operates in the urban area making frequent stops and idling whereas goods movement trucks operate on highways carrying heavy loads moving at a consistent speed with less frequent stops. Selecting vehicles that are tailored for specific duty can ensure lower maintenance costs,

lower fuel consumption, and less downtime.

The real-world data in the wild are clustered in structure most of the time. Most importantly longitudinal clustering is where an individual instance of phenomena has multiple measurements sometimes with hierarchy. For example, there is a natural gas school bus, diesel school bus, and propane school bus which further have data collected from multiple vehicles over time. Each of these can be considered a cluster. The idiosyncrasies of the vocation type, fuel type, and vehicle need to be taken into account while modeling. Developing a global model or individual models per cluster of data is not a feasible solution as the global model cannot learn idiosyncratic about the cluster, and having multiple models per cluster cannot learn across populations resulting in too many models and generalizing well.

To address this issue, a mixed effect model is developed in this chapter where the prior is taken from data with random effects in each cluster. The goal of this study is to develop a single generalized prediction model for estimating the average cost per mile for heavy-duty trucks that have different duty cycles and use different fuel types. This work helps in raising awareness regarding alternative fuel vehicles (AFVs) to the public and their impact on reducing emissions. This would further enable companies to easily analyze the maintenance costs associated and adapt to the AFVs based on the type of vehicle operation, region of operation, and other factors.

6.1 Introduction

Maintenance and Repair (MR) costs play an important role in the total cost of ownership for companies but is not well studied due to limited/lack of data on costs associated with the maintenance of advanced powertrain systems. The maintenance and repair costs reflect the cost of parts and labor for activities such as (i) periodic maintenance activities such as tire rotation, engine oil change, coolant inspection, etc., (ii) corrective maintenance such as replacing failed components such as exhaust system, brakes, transmission, etc. (iii) preventative maintenance such as replacing tires, brakes, etc. before they fail. The important economic indicator for vehicles is the operating cost which includes maintenance and repair costs, fuel costs, and the decreased value of the vehicle over time [153].

Medium and Heavy-Duty vehicles are key to global transportation for activities such as goods

movement, deliveries, services, etc. accounting for about 23% of greenhouse gas (GHG) emissions in the United States [135]. These vehicles are classified into different types based on the maximum loaded weight called gross vehicle weight rating (GVWR). Class 7 heavy-duty vehicles with GVWR 26,001 – 33,000 pounds include furniture trucks, towing trucks, and transit buses whereas class 8 heavy-duty trucks with GVWR of greater than 33,000 pounds include heavy semi-tractors, dump trucks, fire trucks, semi-sleepers, etc. [161]. Heavy-Duty trucks with internal combustion engines (ICE) powered by diesel are predominant resulting in increased pollution, climate change, and health impacts despite the zero-emission and clean air acts [162]. To reduce the emissions from diesel combustion, alternative fuels [163] such as natural gas [164], propane, electric vehicles [165], and hybrid-electric powertrain systems [166] using batteries [167] were introduced as lower emission or zero-emission strategies and have a lower cost of maintenance for the useful life of vehicles [168]. However, studies show that only 6% of alternative fuel vehicles are currently being used by global transportation fleets. The main barrier is promoting the use of AFVs is the public knowledge, opinion, unknown upfront and fuel costs, and vehicle performance [169].

Diesel engine technology has been the preferred powertrain for delivery and goods movement applications. The performance of diesel engines is characterized by high torque and durability that has contributed to it being the preferred technology choice in goods movement applications. However, stringent emissions standards set forth by the US EPA in 2010, meant that diesel engine technology had to embrace a series of after-treatment and emissions control strategy that was aimed at lower tailpipe oxides of nitrogen (NOx) and soot emissions. The introduction of a complex after-treatment system also introduced increased maintenance costs, failure rates, and vocation-specific fuel penalties due to an aggressive emissions control strategy.

Natural Gas Vehicles have emerged as a low oxide of nitrogen (NOx) solution in many of the non-attainment regions of the country. Ultra-low NOx emissions from a three-way catalyst (TWC) based after-treatment system provide a cost-effective solution for emissions control. The low-NOx capability is showcased in urban operations where conventional diesel's emissions control system operation is severely hindered. In a study conducted in Europe, it was found that for long-haul operations a liquified natural gas (LNG) vehicle had a 3% lesser maintenance cost compared to a conventional diesel powertrain. The lower cost of the LNG vehicle was owed to a less complex exhaust after-treatment system. Furthermore, for a straight truck operating in an urban environment

fueled by compressed natural gas (CNG) a 4% higher maintenance and repair cost was estimated. The increased cost was attributed to the up-keeping of the high-pressure CNG fuel system [170].

Like NGVs, propane vehicles operate with a TWC as well as using a stoichiometric platform. Propane vehicles are highly sorted after in school bus applications, with more than 11,000 buses operating throughout the country. Low-fuel cost and abundant fueling infrastructure make propane an attractive alternative fuel option. In comparison to natural gas, propane is used in vehicles with lower GVWR. Propane vehicle GVWR is currently at 33,000 lbs., while heavy-duty natural gas vehicles are operating with a GVWR of up to 56,000 lbs. and can haul a trailer with a combined load of up to 65,000 -70,000 lbs. Although the powertrain technology and emissions control of natural gas and propane are the same and their comparison to a conventional diesel powertrain is on similar metrics, we expect to see significantly different MC estimates between fleets operating propane and NGV.

Electric vehicle (EV) technology has seen significant penetration into the medium and heavy-duty segments in the recent past. The state of California has been pioneering various plugin electric vehicle development for drayage and urban delivery application. The Volvo LIGHTS project in Southern California is a clear example of the state's push toward zero-emissions battery electric trucks in the goods movement sector [171]. California will require large transit agencies in the state to procure 25% of their new vehicles as EVs and by 2040, 100% of their vehicle purchases should be EVs. EVs have been traditionally used in school buses and transit bus applications throughout the country. A report developed by the US Public Interest Research Group (PIRG) suggests that EV transit buses provide combined fuel and MC savings of up to $50,000 a year [172]. Furthermore, the report also cites that electric school buses in California reported an annual savings of $4,400 in MC. The biggest benefit of MC in the EV platform for HD and MD applications is in the brake wear component of the vehicle. Regenerative braking is a significant factor in reducing brake wear. However, the MC related to battery wear has been seldom documented and the changes to this cost as a function of ambient operating temperatures are critical. Extreme hot conditions with insufficient cooling and extremely cold conditions can cause significant stress to the batteries and thereby deteriorating health. Furthermore, EVs have far fewer components than an internal combustion engine powertrain. However, the deterioration of high-power electronics and the associated downtime related to these failures could be a direct function of the aggressiveness

of day-to-day activity.

Real-world data collected by WVU show that urban operation adversely affects fuel consumption and the performance of modern diesel technology. Particularly, the diesel particulate filter (DPF) undergoes frequent regeneration due to excessive soot loading, and the cost of DPF ash cleaning is seldom considered during the maintenance cost evaluation. AFV vehicles operating on both natural gas and propane will not be subjected to the maintenance cost of complex particulate matter (PM) and NOx after-treatment systems. The exhaust gas recirculation (EGR) systems of vehicles operating in urban scenarios are significantly dirtier than long-haul operations and consequently, EGR cooler maintenance costs of diesel vehicles operating in urban applications are high. The proposed study will be analyzing the comparative maintenance cost at a fine component level resolution, wherein, every component that could potentially be affected due to the duty cycle will be considered as a key parameter in maintenance and repair cost.

Modern diesel vehicles have a suite of onboard sensors that are used for onboard diagnostic purposes. Many of these sensors are subject to harsh operating conditions and frequent failures whereas AFV control systems typically operate with minimal sensor-based diagnostics reducing failure rates and maintenance costs (MC). The cost of consumables in a diesel vehicle is higher than that of AFVs due to the use of diesel exhaust fluid (DEF) for NOx control. Finally, the brake maintenance cost of electric vehicles would be the lowest due to the availability of regenerative braking. Although HD diesel has advanced engine brakes that improve the life of wheel brakes, EVs operating in transit buses, school buses, and medium-duty applications will have lower brake maintenance than comparable diesel and other AFVs as well. The maintenance and repair cost data is generally collected from fleet owners/operators or by surveying truck original equipment manufacturers (OEMs) [173].

Based on the technology and truck type, the patterns of maintenance differ impacting the maintenance and repair costs. Maintenance and Repair costs for heavy-duty diesel trucks were well studied and documented based on literature sources [174] but there are very few studies related to alternative fuel heavy-duty trucks. A detailed component level MR cost and the total cost of ownership for different electrified propulsions is studied on German cars which tend to have lower MR cost due to less complex power train [175]. The TCO associated with class 8 long-haul, class 8 short-haul, and class 4 parcel delivery vehicles using 6 power trains such as conventional

(diesel), hybrid-electric, plug-in hybrid electric, compressed natural gas, fuel cell electric, and battery electric are analyzed [176]. The total cost of ownership for internal combustion, hybrid, and electric light-duty vehicles was studied taking vehicle components into consideration based on a few assumptions [177]. A study was performed to promote the diffusion of battery electric vehicles (BEV) based on the total cost of ownership (TCO) [178]. The maintenance cost for transit buses involving various maintenance costs was studied. Theoretical frameworks have been developed for promoting the diffusion of alternative fuel vehicles [157].

Despite these studies, the adoption of alternative fuel vehicles has not reached the mark expected. Most of the studies till now are relied on simulations or theoretical frameworks taking assumptions or a vehicle type into account. Development of a model based on the historical maintenance and repair data to estimate the maintenance cost would be a feasible solution that can be used by consumers.

With Industry 4.0 technologies, Artificial Intelligence has gained success in a wide range of applications in the automotive and transportation sector. These technologies enable the collection of vast data and making use of the data for studying difficult tasks or time-consuming to perform using existing methods. Studies applied machine learning algorithms for predictive maintenance [93], failure of components in trucks [179], estimating remaining useful life [145]. Studies using analytical models are used to estimate the maintenance and repair costs at the component level for heavy-duty trucks using battery electric and fuel cells.

Most of the studies are analytics based on an individual vehicle type or just taking the change in cost for various components. The duty cycle, region of operation, season, and frequency of maintenance greatly affect the maintenance cost impacting the total cost of ownership. Taking these factors into consideration while modeling the maintenance and repair costs is important. However, there is very little data on maintenance and repair costs for heavy-duty vehicles using alternative fuels such as compressed natural gas, liquefied propane gas, battery electric, and hydrogen fuel cells, while data related to conventional diesel trucks are available. But the maintenance needs and patterns in different trucks with varying technologies differ widely, and it cannot be assumed that conventional diesel truck MR costs can be applied to represent other alternative fuel heavy-duty trucks.

This study intends to bridge the gap in research by developing a generalized machine learning

model that can be used to estimate the average cost per mile ($/mile) for various medium- and heavy-duty trucks using diesel or an alternative fuel such as natural gas or propane. This could be an important contribution as the model can be used by the consumer to estimate the maintenance and repair costs given the activity being performed by the truck, the region of operation, the mileage that is expected to hit, and the fuel type enabling procurement decisions. Most importantly, this study uses historical maintenance and repair data collected from heavy-duty vehicles such as delivery trucks, goods movement, school buses, transit buses, refuse trucks and vocational trucks operated using diesel or natural gas, or propane gas. It is important to consider the activity, fuel type, region of operation, mileage, etc since this could be the basis for new and improved public policy and marketing for alternative fuel vehicles. The prediction model helps estimate the average cost per mile throughout operation for a vehicle emphasizing factors that influence the choice of vehicle, fuel type, and the total cost of ownership (TCO). The input features remained the same as discussed in section 5.2.1 with few other features representing the activity performed by the vehicle. Data has been collected from multiple vehicles where each vehicle underwent a number of periodic, preventative, and corrective maintenance. The maintenance records include cascaded maintenance where preventative or corrective maintenance is identified during the periodic maintenance or multiple fixings/replacements are performed during a single maintenance.

6.2 Methodology

The duty cycle of heavy-duty vehicles affects the maintenance cost associated. This study utilizes a large volume of data collected from different medium- and heavy-duty trucks using different fuel types such as diesel, natural gas, propane, and electric vehicles with different duty cycles performing activities such as goods movement, delivery, school bus, refuse, and vocational. The data collection has been performed following the technical proposal by WVU CAFEE and the Department of energy (DOE). The data is collected in partnership with Clean City Coalitions to reach fleet companies that operate AFVs. The distribution of data grouped by fuel type and activity is shown in Figure 6.1.

The data has been pre-processed to remove duplicate records and missing values. A Z-score method is used to calculate the interquartile range (IQR) for the target variable. Based on the box

Table 6.1: Features and Description

Feature	**Description**	**Feature**	**Description**
UnitNumber	Unique Identification number per vehicle	Vehicle Type	Activity type performed by the vehicle
Date	Date of maintenance or repair performed	TotalCost	Total cost for the maintenance or repair performed
Mileage	Mileage recorded on the vehicle	TBM	Number of days from previous maintenance
FuelType	Type of fuel	VAge	Age of the vehicle
PartofTruck	Part of the truck on which maintenance or repair performed	MilesPerDay	The number of miles recorded per day (calculated based on previous mileage)
Comments	Type of Maintenance	MaintenanceYear	Calendar year in which maintenance is performed
Engine	Engine Model	MaintenanceNum	Count of the maintenance recorded on the vehicle
EngineYear	Year of engine manufacture	Season	The season during which the maintenance or repair occurred
Make	Make of vehicle	RCI	Road congestion index for the region
Region	Region of operation of the vehicle	CostPerMile	Cost Per Mile

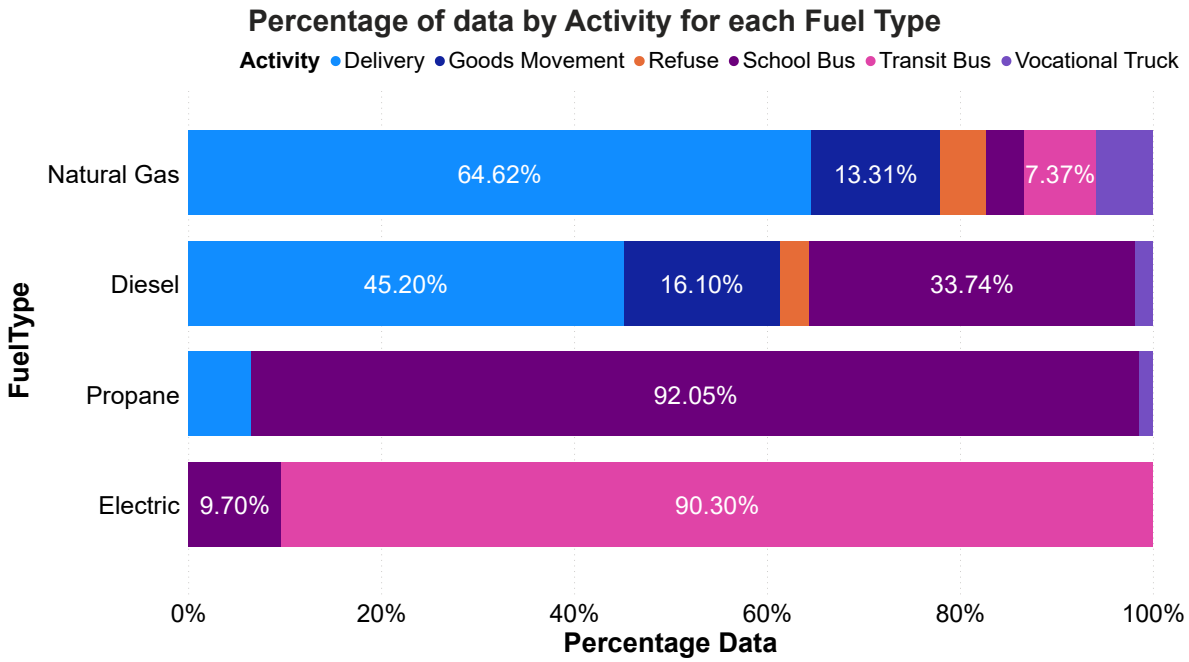

Figure 6.1: The distribution of data based on activity and fuel type

plots for IQR, a careful inspection of data points outside the whiskers is performed to determine whether the data points should be considered an outlier. The feature correlation is then performed on the pre-processed data to determine the association between features. The feature association matrix for the input features and the target variable is shown in Figure 6.2. The pre-processing involves the same steps discussed in section 5.2.3 on features mentioned in Table 6.1.

The maintenance and repair data are categorized into 3 maintenance types: periodic, preventative, and corrective with 5 major parts of trucks: chassis, engine and transmission, exhaust and emission, fuel system, tire & brake.

1. Periodic maintenance is planned or regularly scheduled maintenance such as engine oil change, tire rotation, engine inspection, and other routine work. These are based on manufacturer recommendations or mandated by law.

2. Preventative maintenance is performed proactively to reduce vehicle downtime and to prolong the useful life of components. These include checking for issues/failure in major components.

3. Corrective maintenance is a significant portion of overall maintenance and repair costs. These

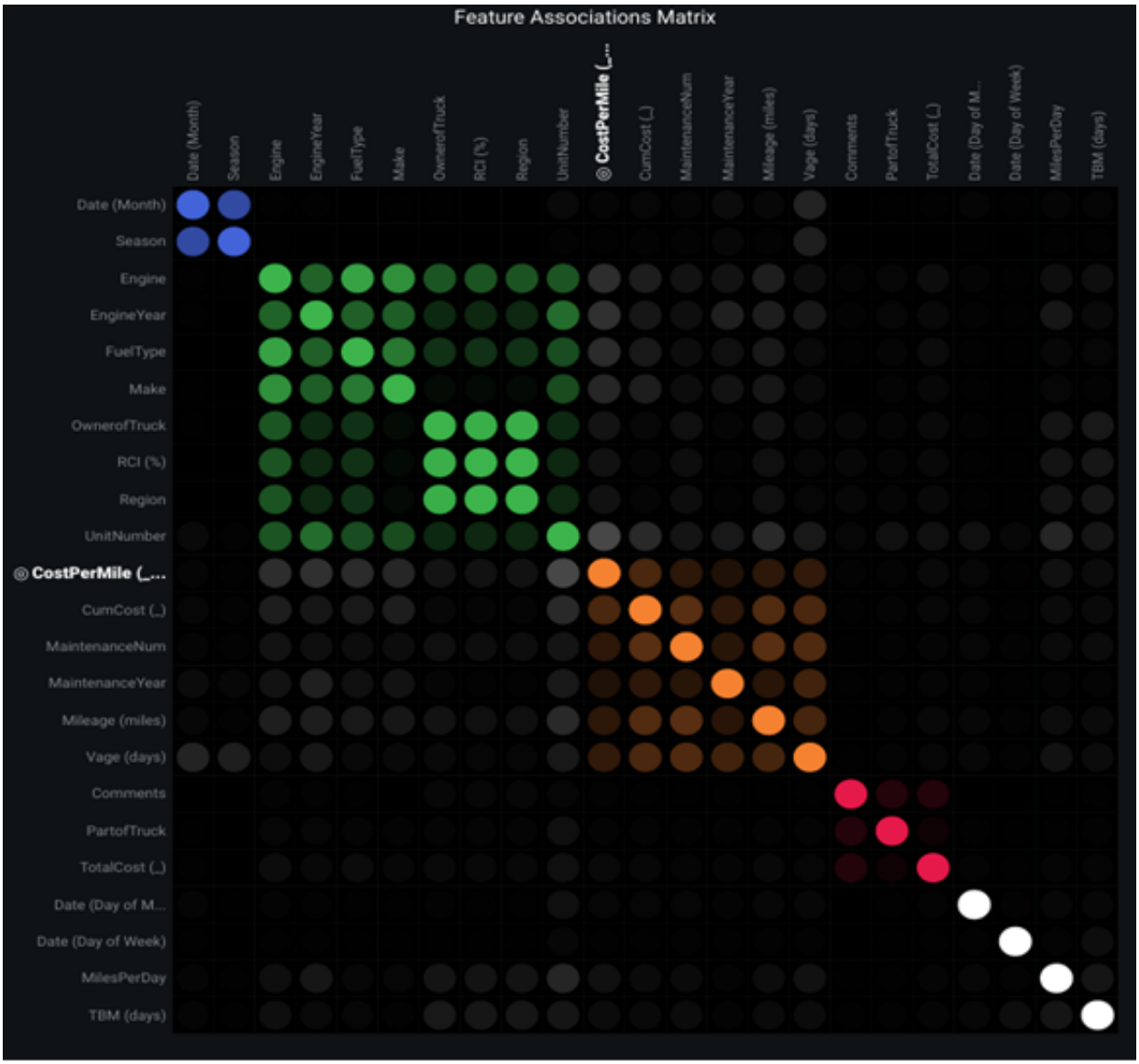

Figure 6.2: Feature Association Matrix

are major fixings such as the replacement of transmission, and fuel system which cost significantly over the lifetime of the vehicle.

The total cost for maintenance and repair includes the cost of parts and the labor cost. Hence the cost-per-mile is calculated as

$$\text{Cost Per Mile} = \frac{\text{Part Cost} + \text{Labor Cost}}{\text{Mileage}} \tag{6.1}$$

The data being used for this analysis are clustered longitudinally, where random effects exist between subgroups of vehicle types and fuel types. Furthermore, the maintenance of each vehicle is performed at different intervals of time varying in the total miles of operation for vehicles. For example, a few fleet managers might perform more periodic maintenance whereas a few companies might not perform regular maintenance resulting in more corrective maintenance. Hence the time interval between maintenance for a given vehicle is not regular resulting in repeated or longitudinal clustered data. By carefully analyzing the data, the fixed and random effect features are identified. The features varying within a cluster such as Mileage, TBM, VAge, and MilesPerDay are identified to be random effect features whereas MaintenanceYear and Region are considered as both fixed and random effect features as they tend to have a constant value for some clusters. Along with the input features a cluster id is passed as input for each cluster.

6.2.1 Mixed Effect Models

In this chapter, the goal is to develop a generalized model to capture the mixed effects in the data. Most of the machine learning algorithms assume the training data to be i.i.d. which is commonly violated in longitudinal data where there is a high correlation between subgroups. The vocation-based data analysis and graphical analysis have been performed previously on the data collected using excel and MATLAB [180]. The distribution of residual and random effects is of importance for the accurate and unbiased estimation of the model. The fixed-effect model works well when studies include analysis of identical data and the goal is to model the identified population rather than generalizing for other populations. On the other hand, the random effects are performed on data from a series of experiments where the subjects differ impacting the results to generalize well for various scenarios [181]. Mixed-effects regression models estimate fixed and random effects in a single mode.

Mixed effects regression models are used to model data that has group-level and global trends in data. Linear Mixed Models (LMM) are extended linear models introduced to capture the dependencies using random effects and effects between covariates and using fixed effects with correlated multilevel longitudinal data [182]. The typical linear model is represented as $y = X\beta + \epsilon$ where $X\beta$ represents fixed term and ϵ represents error. An incorrect specification of random effects in linear models has consequences on the maximum likelihood estimator [183]. In longitudinal clustered data, the variability within the group or between groups affects the outcome. One way of handling such data is to aggregate the individual group data, which then becomes independent. However, this approach does not consider all data missing the key patterns within the group. Another approach analyzes each group at a time resulting in an individual model for each group. However, this approach does not take information from the global population. LMMs are in between making the trade of the two alternative approaches. The key idea of the mixed model is to take into consideration both the fixed and random effects. Fixed effects are a parameter that is fixed without variation whereas random effects are parameters that are random variables like linear regression, but the parameters are fixed. The true population β is a random normal variable with mean μ and standard deviation σ given by $\beta \sim N(\mu, \sigma)$. The linear mixed models are represented as

$$y_i = X_i\beta + b_i Z_i + \epsilon_i, \tag{6.2}$$

where y_i is $n_i \times 1$ vector containing responses for n_i observations in cluster i, X_i is $n_i \times p$ matrix of fixed-effects covariates, b_i is $q \times 1$ unknown vector of random effects for the cluster i, Z_i is $n_i \times q$ matrix of random-effects covariates, ϵ_i is $n_i \times 1$ vector of errors and β is an unknown vector of fixed effects coefficients. The random part, $Z_i b_i$, is assumed linear. b_i and ϵ_i are assumed to be independent and identically distributed normal random vectors given by $b_i \sim N(0, D)$ and $\epsilon_i \sim N(0, R_i)$, where D and R_i are covariance matrices for b_i and ϵ_i, respectively.

In this chapter mixed-model regression analysis is performed which deals with longitudinal data having within and between group variances. The approach includes both fixed effects which define overall change over time and random effects accounting for variability among clusters. Expectation Maximization (EM) algorithm is used to iteratively learn the maximum likelihood and the random effect coefficients. However, this requires the functional form to be specified which is difficult, especially for complex longitudinal clustering. To address this using a tree-based method, a semi-

parametric mixed model with fixed effects non-parametric tree model, and random effects part is proposed [184]. A similar approach named mixed effect regression trees (MERT) is proposed in [185].

Random forests are an ensemble model that combines individual decision trees [186] to improve the predictive capability of the model and reduce the variance [187]. The idea of bagging [188] is applied to the random forest for bootstrap aggregation on de-correlated trees controlled by several trees and the number of variables per split. Each de-correlated tree in the forest aims at minimizing the prediction mean square error (MSE) resulting in the random forest regression function minimizing the point-wise mean square error (MSE). For a regression model, the minimization of squared error loss is the conditional mean of the target variable given the data. However, while using the random forest model observations are assumed independent and ignore the underlying assumptions such as linearity and distribution of data. Ignoring the correlation in data results in lower pointwise predictions. Therefore, the fixed effects part of MERT was replaced with random forests to develop mixed effects random forests (MERF) [189]. The advantages of random forest and linear random effects have been combined to develop a mixed-effects random forest model given by the form:

$$y_i = f(X_i) + Z_i b_i + \epsilon_i, \tag{6.3}$$

$$b_i \sim N(0, D) \text{ and } \epsilon_i \sim N(0, R_i), \quad i = 1, 2, \ldots, n,$$

where y_i is $n_i \times 1$ vector containing responses for n_i observations in cluster i, X_i is $n_i \times p$ matrix of fixed-effects covariates, b_i is $q \times 1$ unknown vector of random effects for the cluster i, Z_i is $n_i \times q$ matrix of random-effects covariates, ϵ_i is $n_i \times 1$ vector of errors and β is an unknown vector of fixed effects coefficients. The random part, $Z_i b_i$, is assumed linear. b_i and ϵ_i are assumed to be independent and identically normal distributions given by $b_i \sim N(0, D)$ and $\epsilon_i \sim N(0, R_i)$, D and R_i are covariance matrices for b_i and ϵ_i respectively. The non-linear function $f(X_i)$ learned using random forest is used to represent the fixed effect in MERF. An expectation-maximization (EM) algorithm [190, 191] is used to iteratively fit the MERF by optimizing one parameter while keeping others fixed until convergence is reached. The EM algorithm for fitting MERF is as follows:

1. Start with default values for variance (σ_i), random effects coefficient ($\hat{b}_i$) and the diagonal matrix of unknown variance ($\hat{D}$).
2. Calculate the response variable ($y^*_{i(r)}$), the estimated function ($\hat{f}(x_{ij})$), and the random effects coefficient ($\hat{b}_{i(r)}$).
   - Calculate response variable, $y^*_{i(r)} = y_i - Z_i\hat{b}_i$.
   - Estimate fixed effects by taking bootstrap samples (y^*_{ij}, x_{ij}) using the random forest.
   - Find the random effects coefficient $\hat{b}_i$ at cluster i using the estimated $\hat{f}(x_{ij})$ from random forest.
3. Compute variance $\hat{\sigma}^2$ and $\hat{D}$ from estimated residuals and random effects respectively.
4. Repeat step 2 and 3 until convergence.

The convergence of the MERF algorithm is monitored using generalized log-likelihood (GLL) given by:

$$GLL(f, b_i|y) = \sum_{i=1}^{n} \Big\{ [y_i - f(X_i) - Z_i b_i]^T R_i^{-1} [y_i - f(X_i) - Z_i b_i] + b_i^T D^{-1} b_i + \log|D| + \log|R_i| \Big\}, \tag{6.4}$$

where R_i is the covariance matrix for ϵ_i and D is the covariance matrix for b_i.

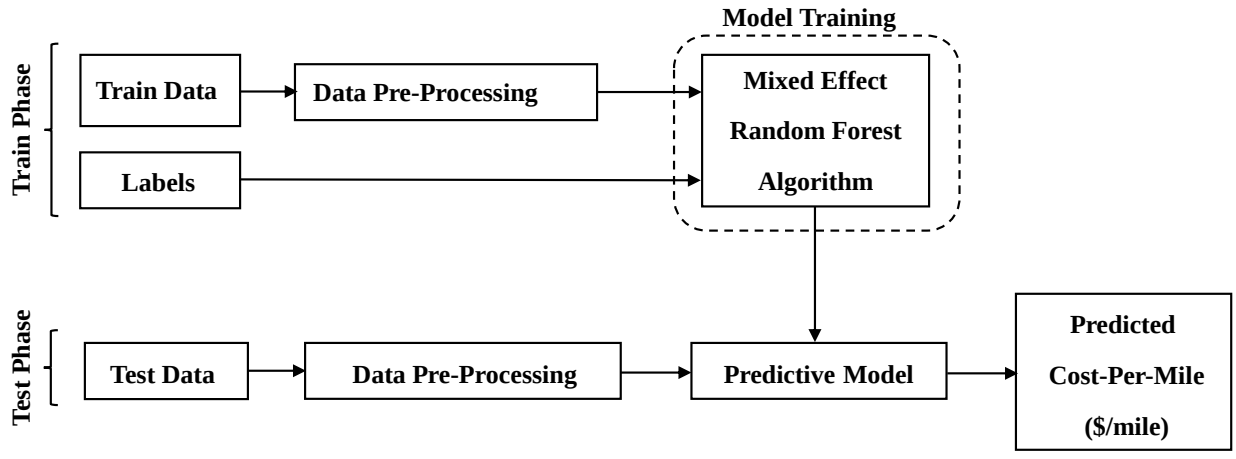

Figure 6.3: Workflow for MERF Model

Table 6.2: The number of vehicles available per cluster

Vehicle Type	CV MAE Diesel	Natural Gas	Propane
School Bus	27	15	81
Delivery	24	30	2
Vocational	4	8	2
Refuse	5	6	0
Transit Bus	0	2	0
Goods Movement	14	48	0

The MERF assumes the random effects term to be correct for estimating the forest function and assumes out-of-bag predictions from the forest to be correct for estimating the random effect's part [192]. The unused observations from the forests sub-tree are used in Out-of-Bag predictions [187].

A typical workflow for MERF model is shown in Figure 6.3. Once the model is fitted, it can be used to make predictions on known clusters as well as new clusters that are not seen during training. For the known cluster data, the predictions are given by: $\hat{y} = f(X) + b_i Z$ whereas for new clusters the predictions only include fixed effect given by $\hat{y} = f(X)$.

6.3 Results and Discussion

The qualitative data collected from different fleet management companies are used for modeling a MERF model. The number of vehicles per activity per fuel type in the data collected is shown in Table 6.2.

The entire dataset is divided into the train, validation, and test datasets. Data related to one vehicle per cluster is randomly selected based on the cluster id to form a test dataset. This test dataset is not seen by the model during the training process and is used to evaluate the model performance once the model is trained completely. To test the model on a new cluster that has not been presented for training, transit bus data are used. There are only two transit buses with very few data points, hence the transit bus cluster has not been used for training. The remaining data is split into 70% and 30% randomly based on the vehicle unit number to form train and validation datasets respectively. The training dataset is used to fit the model and the validation

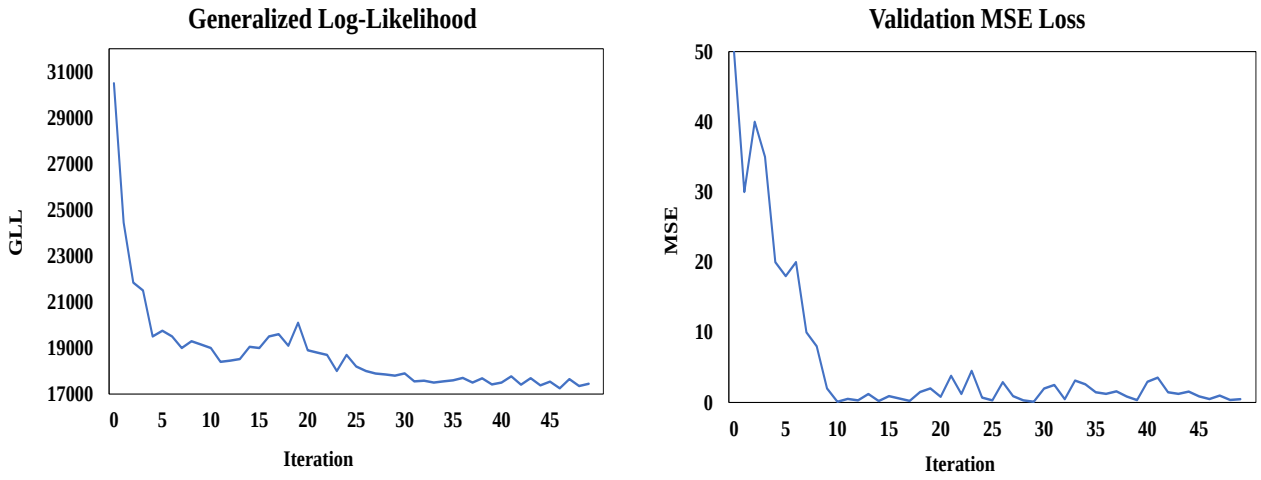

Figure 6.4: Generalized Log-Likelihood (Left Panel) and Mean Square Error (Right Panel)

dataset is used to evaluate the model performance at every iteration. The cost-per-mile continuous feature is the target variable with the remaining features being input variables. The MERF model is trained for 50 iterations with the number of trees being 50 in the random forest. The generalized log-likelihood during the training of the MERF model over each iteration is shown in Figure 6.4. From the plot, the model converges by the 50th iteration.

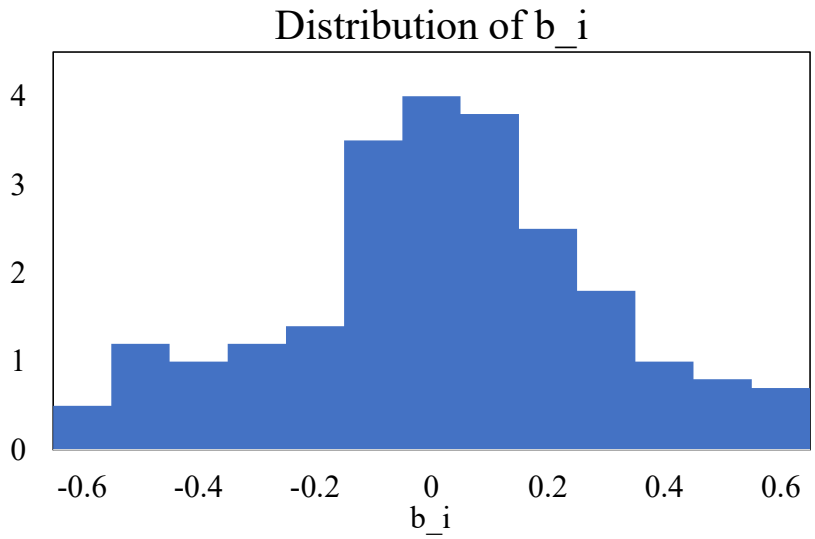

Figure 6.5: Distribution of values of b_i after training

Along with the GLL, the validation mean square error is shown in Figure 6.4. The model

Table 6.3: Evaluation Metrics on Test Data Clusters

Fuel Type - Vehicle Type	R^2(%)	**MAE ($/mile)**	**MSE ($/mile)**
Diesel - School Bus	79.81	0.0351	0.00205
Natural Gas - School Bus	90.32	0.0284	0.00144
Propane - School Bus	76.98	0.0393	0.00434
Diesel – Delivery Truck	85.26	0.0486	0.00351
Natural Gas – Delivery Truck	75.89	0.0208	0.00077
Propane – Delivery Truck	81.63	0.0188	0.00058
Diesel – Vocational Truck	95.92	0.0136	0.00355
Natural Gas – Vocational Truck	96.98	0.0459	0.00632
Propane – Vocational Truck	65.89	0.1011	0.01768
Diesel – Refuse Trucks	76.47	0.1710	0.10646
Natural Gas – Refuse Trucks	94.55	0.0228	0.01103
Diesel–Goods Movement	69.82	0.1395	0.00292
Natural Gas – Goods Movement	96.87	0.0651	0.00736
Natural Gas – Transit Bus	95.46	0.0114	0.00086

achieved a coefficient of determination (R^2) of 98.96% and a mean square error (MSE) of 0.0089 $/mile for the training dataset and R^2 of 94.31% and MSE of 0.0312 $/mile for the validation dataset. The trained model also holds the distribution of b_is learned over iterations. The b_i is different for each cluster but is drawn from the prior data distribution. The distribution of learned b_is is shown in Figure 6.5.

Once the model is trained, the random forest model $f(X)$ along with learned b_i is used to predict the cost-per-mile for the unseen test dataset. Sample data from each cluster are used to test the model performance. The results indicate that the MERF model generalizes well for the unseen clustered test dataset. For a few clusters such as diesel – school bus, and propane – vocational the model performs reasonably well due to very few data points and large variations of data within the cluster. To test the performance on clusters that are not available during the training, the natural gas transit bus data is used. The values of performance metrics including coefficient of determination (R^2), mean absolute error (MAE), and mean square error (MSE) for each cluster in the test dataset are presented in Table 6.3.

Table 6.4: School Bus Test Data Summary

Fuel Type	**Diesel**	**Natural Gas**	**Propane**
Maintenance Year	2015-2019	2018-2022	2013-2022
Mileage (1000 miles)	5-130	101-160	2-93
Number of Periodic Maintenances	30	3	40
Number of Preventative Maintenances	3	27	4
Number of Corrective Maintenances	1	8	27

6.3.1 School Bus

The summary of test data for a diesel and natural gas school bus operated in California and a propane school bus operated in Colorado is presented in Table 6.4. The diesel school bus had more periodic maintenance performed with few preventative maintenances and one corrective maintenance related to the engine and transmission performed. The natural gas school bus vehicle has recorded preventative and corrective maintenance related to chassis, engine & transmission, exhaust & aftertreatment, and a few periodic maintenance. The engine & transmission corrective maintenance in the natural gas vehicle has incurred high maintenance costs after 125,000 mileage. The propane school bus underwent more corrective maintenance and preventative maintenance related to the engine and transmission after 30,000 miles with frequent periodic maintenance. The corrective and periodic maintenance of the engine and transmission involved higher maintenance costs but lower than the cost incurred for tire and break.

The comparison of average cost per mile over the duration for diesel, natural gas, and propane school bus data is shown in Figure 6.6. The diesel vehicle showed the highest average cost per mile with just one corrective maintenance related to the engine & transmission at around 70,000 miles as the periodic maintenance of the engine & transmission, fuel system was costly. Natural gas vehicles showed an increasing trend in the average cost per mile as the duration of operation increased. However, the observed values are for mileage ranging from 101,000 – 160,000 miles. Though propane vehicles underwent several preventative and corrective maintenance before hitting 52,000 miles, the average cost-per-mile is observed to be lower than diesel and natural gas vehicles.

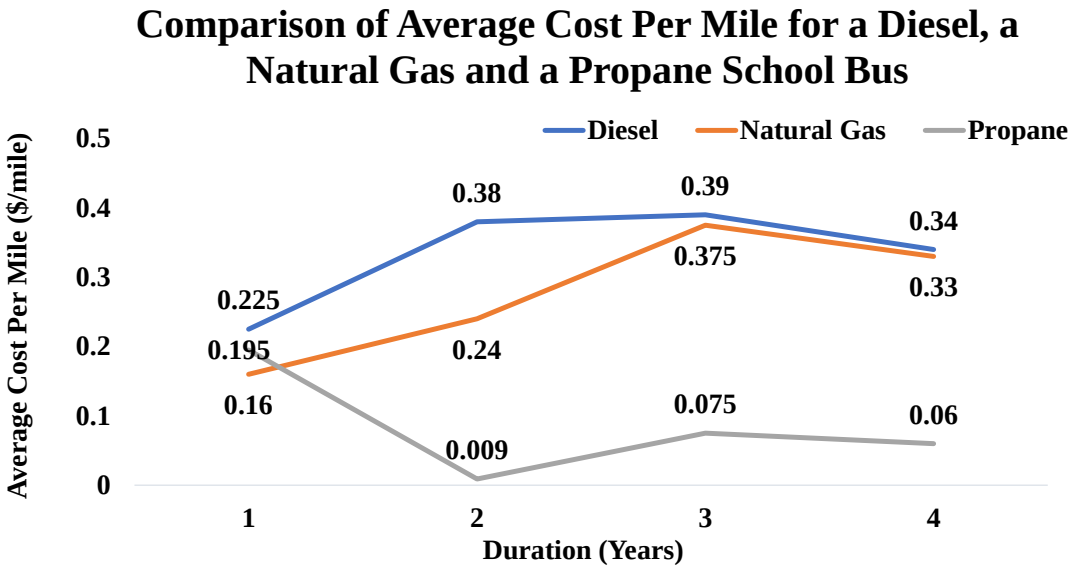

Figure 6.6: Average predicted cost per mile in school bus

6.3.2 Delivery Truck

The summary of test data for a diesel and propane delivery truck operated in South Carolina and a natural gas delivery truck operated in Pennsylvania is presented in Table 6.5. The diesel delivery truck underwent a similar number of periodic and corrective maintenance with a couple of preventative maintenance performed throughout the operation. The corrective maintenance on the chassis had a total cost much higher after the vehicle recorded mileage greater than 104,000 miles with one major engine & transmission at a mileage of around 68,000. Engine & transmission

Table 6.5: Delivery Test Data Summary

Fuel Type	Diesel	Natural Gas	Propane
Maintenance Year	2012-2017	2017-2021	2015-2022
Mileage (1000 miles)	12-106	0.4-462	6.5-100
Number of Periodic Maintenances	13	40	21
Number of Preventative Maintenances	2	17	5
Number of Corrective Maintenances	10	55	15

preventative maintenance has a total cost of 10 times more than tire & brake preventative maintenance. The cost of periodic maintenance for tire & brake is 2 to 3 times higher than chassis whereas the periodic maintenance cost for engine & transmission is lower than all. Since the natural gas delivery truck has operated for a very high mileage of 460,000, it has recorded a high number of corrective maintenance and periodic maintenance with few preventative maintenance. The chassis has less periodic maintenance costs followed by engine & transmission and then tire & brake. But the preventative maintenance costs for engine & transmission are higher than tire & brake. The fuel system has the lowest corrective maintenance costs among all parts. The tire & brake components have seen corrective maintenance costs more than double after 113,000 miles whereas exhaust & emissions corrective maintenance at 160,000 miles is higher than the cost above 330,000 miles. The engine & transmission went through much corrective maintenance with one maintenance at

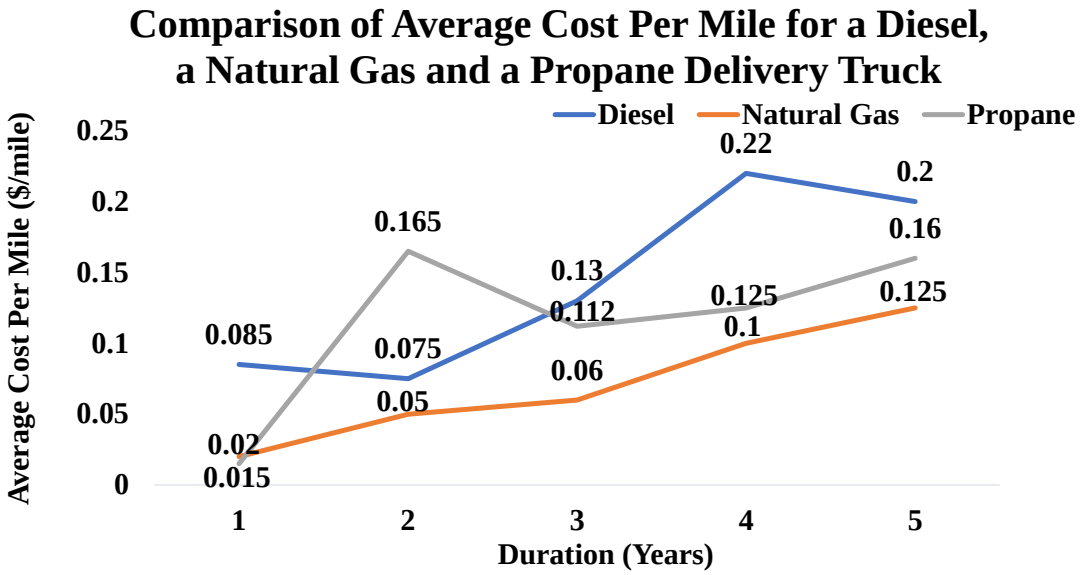

Figure 6.7: Average predicted cost per mile in delivery truck

360k miles incurring 10 times higher cost than the maximum of other corrective maintenance costs on the same part. The corrective maintenance cost for the chassis has been observed to be more than doubled every 100,000 miles. The periodic maintenance for the chassis had constant costs

throughout the operations. For tire & brake, the periodic maintenance costs have increased and for engine & transmission, the mileage below 13,000 and above 74,000 had a higher value. One-third of the corrective maintenance for the engine & transmission had higher costs at various mileages. The preventative maintenance for this vehicle was only performed for the engine & transmission with increasing maintenance costs.

The comparison of average cost per mile over the duration for diesel, natural gas, and propane delivery data is shown in Figure 6.7. The natural gas delivery truck showed the lower average cost per mile even with the highest mileage accumulated in 5 years. The trend shows an increase in the value throughout the operation. A similar increasing trend is observed for diesel delivery trucks but with a higher average cost per mile. The propane vehicle projected nearly similar average cost per mile every year except the first year.

6.3.3 Vocational Truck

The summary of test data for a diesel vocational truck operated in Ohio, a natural gas vocational truck operated in California, and a propane vocational truck operated in Rhode Island is presented in Table 6.6. The diesel vocational truck recorded a high number of corrective and preventative maintenance related to chassis and engine & transmission for a total of 6000 miles over 2 years. The corrective maintenance cost for the engine & transmission was very high at 3000 miles whereas for the chassis both the corrective maintenance and preventative maintenance cost had a fluctuating trend. The natural gas vocational truck for 3 years had frequent periodic maintenance related to the chassis with corrective maintenance related to the chassis and fuel system. The corrective maintenance costs are much higher compared to periodic and preventative maintenance costs. The propane vocational had periodic maintenance every year with one corrective maintenance related to the engine & transmission which has 3.5-4 times lower maintenance cost than periodic maintenance.

The comparison of average cost per mile over the duration for diesel, natural gas, and propane vocational truck data is shown in Figure 6.8. The average cost-per-mile for a diesel vehicle is recorded as very high for a mileage range of 1000-6000 miles with an increasing trend. A similar increasing trend is observed in a natural gas vocational truck for mileage range 5000-9000. The propane vocation trucks have lower average cost-per-mile even with higher mileage of 13,500-50,000 miles.

Table 6.6: Vocational Test Data Summary

Fuel Type	Diesel	Natural Gas	Propane
Maintenance Year	2021-2022	2020-2022	2016-2019
Mileage (1000 miles)	1-6	5-9	13-50
Number of Periodic Maintenances	1	10	5
Number of Preventative Maintenances	4	1	0
Number of Corrective Maintenances	8	2	1

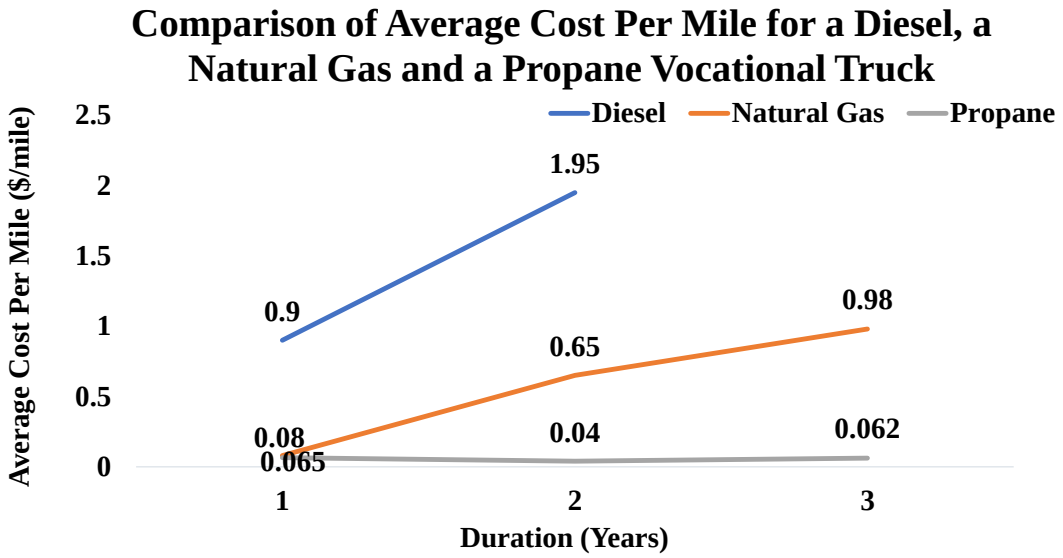

Figure 6.8: Average predicted cost per mile in vocational truck

6.3.4 Refuse Truck

The summary of test data for a diesel refuse truck operated in Ohio and a natural gas refuse truck operated in California is presented in Table 6.7. The diesel refuse truck went through 3 corrective maintenance related to chassis, tire & brake before the truck records 7913 miles with one preventative maintenance at 2750 miles. The natural gas refuse truck had many corrective maintenances related to chassis, tire & brake, and engine & transmission most of the corrective maintenances were observed during periodic and preventative maintenance. The maintenance cost for all types of maintenance had a fluctuating total cost with a few higher costs at some mileage.

The comparison of average cost per mile over the duration for diesel and natural gas refuse

Table 6.7: Refuse Test Data Summary

Fuel Type	Diesel	Natural Gas
Maintenance Year	2021-2022	2020-2022
Mileage (1000 miles)	2.7-7.9	42-83
Number of Periodic Maintenances	3	21
Number of Preventative Maintenances	0	22
Number of Corrective Maintenances	3	66

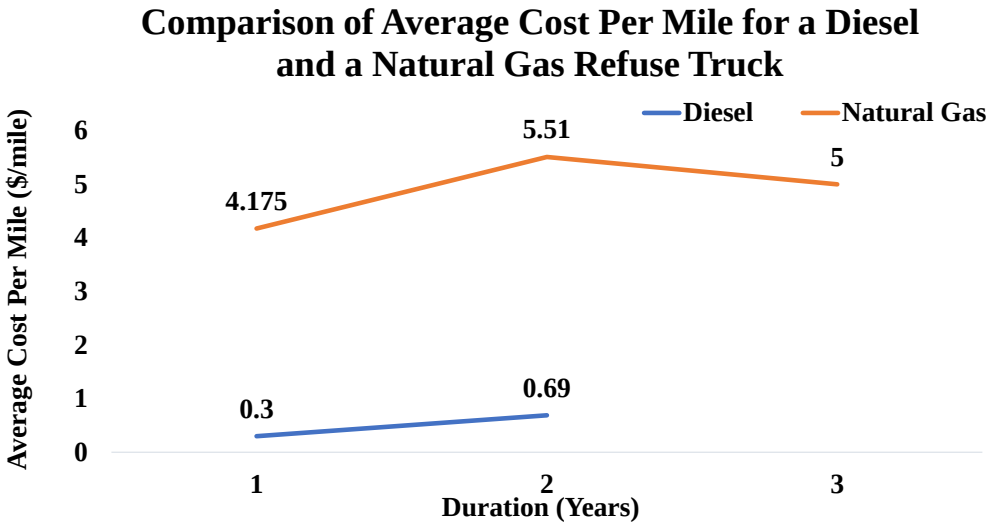

Figure 6.9: Average predicted cost per mile in refuse truck

truck data is shown in Figure 6.9. The diesel refuse has less average cost-per-mile with mileage ranging from 5000-8000 miles with an increase in value over the years. However, the estimated values for natural gas are very high as the vehicle has data with mileage greater than 42000 miles.

6.3.5 Goods Movement Truck

The summary of test data for a diesel and a natural gas goods movement truck operated in California is presented in Table 6.8. The diesel goods movement truck has corrective maintenance related to the engine & transmission, chassis, and exhaust & emissions after recording 114,000 miles observed during regular periodic maintenance. Like, refuse trucks, the natural gas goods movement truck also had many corrective maintenances related to chassis, tire & brake, engine

Table 6.8: Goods Movement Truck Test Data Summary

Fuel Type	Diesel	Natural Gas
Maintenance Year	2010-2014	2017-2021
Mileage (1000 miles)	32-296	30-153
Number of Periodic Maintenances	15	15
Number of Preventative Maintenances	1	11
Number of Corrective Maintenances	8	25

& transmission, and fuel system. The number of corrective maintenance on this vehicle is almost equal to the number of periodic and preventative maintenance together.

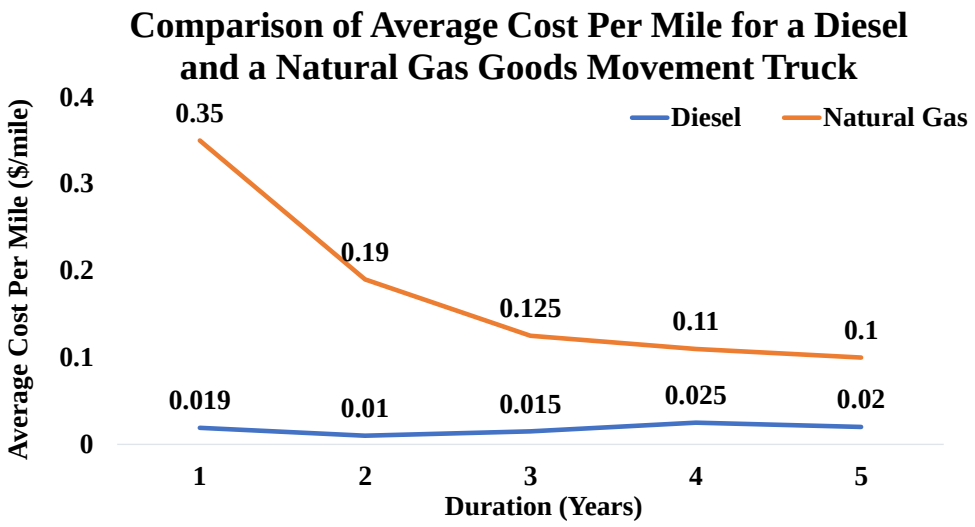

Figure 6.10: Average predicted cost per mile in goods movement truck

The comparison of average cost per mile over the duration for diesel and natural gas goods movement truck data is shown in Figure 6.10. The natural gas goods movement truck had decreasing average cost-per-mile throughout the operation as mileage increased from 30,000 to 152,000 whereas the diesel goods movement truck had a similar value throughout the operation with mileage of 32,700 – 296,000.

Table 6.9: Transit Bus Test Data Summary

Fuel Type	**Natural Gas**
Maintenance Year	2016-2022
Mileage (1000 miles)	2.1-311
Number of Periodic Maintenances	99
Number of Preventative Maintenances	104
Number of Corrective Maintenances	29

6.3.6 Transit Bus

Since there is no data for transit buses using diesel and propane fuels, the transit bus data is not presented for training the model during the training or validation phases. The summary of maintenance data for a natural gas transit bus operated in California is presented in Table 6.9. The natural gas transit bus has operated for 310,000 miles over 6 years. The vehicle had comparably a smaller number of corrective maintenance due to frequent periodic and preventative maintenance performed. The total cost for periodic maintenance of the chassis was higher than the cost for the engine & transmission and tire & brake. The average corrective cost for the engine & transmission is almost equal to the average corrective cost for the tire & brake and half the corrective cost for the fuel system.

The predicted average cost-per-mile for this vehicle is compared with the average cost-per-mile for the original test data as shown in Figure 6.11. This shows how well the model is generalized to unseen cluster data and the performance of the model. The higher average cost-per-mile in the initial year of operation is due to the replacement of the fuel system. The vehicle has seen an increasing trend in the average cost-per-mile for maintenance done.

Overall, most corrective maintenance has been identified during periodic or preventative maintenance. The total cost incurred for maintenance based on the maintenance type, part of the truck, etc. may not always have a trend but the average cost-per-mile calculated using the total cost and the mileage of the vehicle gives interesting insights for each of the truck types. For example, the natural gas delivery truck having many corrective maintenance has less cost-per-mile whereas the natural gas refuse truck went through a higher number of corrective maintenance resulting in a higher average cost-per-mile. From the comparison plots, it is observed that the school buses

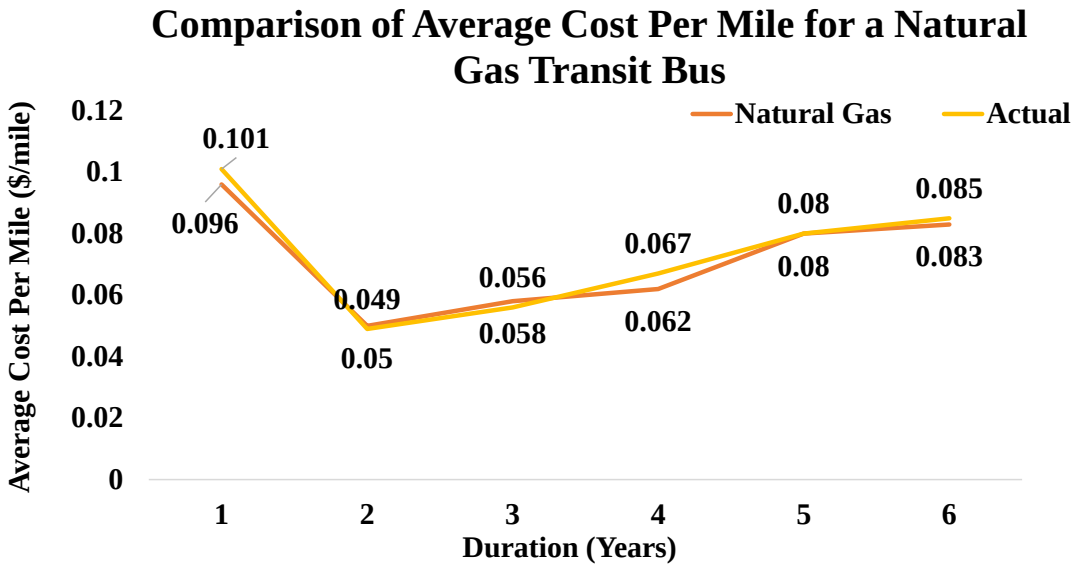

Figure 6.11: Average predicted cost per mile in transit bus

and vocational trucks which have stop-and-go activity operated at lower speeds have lower average cost-per-mile using propane fuel compared to natural gas and diesel whereas the delivery trucks have lower average cost-per-mile using natural gas. The refuse and goods movement trucks using diesel fuel have lower average cost-per-mile. The variation is expected as modern diesel vehicles are designed to have better fuel efficiency with large weights at higher speeds operated in long routes compared to lower speeds with short distances. Hence the goods movement trucks have lower cost-per-mile with diesel than other vehicles such as school buses, delivery trucks, etc.

6.4 Conclusion

Maintenance cost is considered an important factor for the total cost of ownership while purchasing a vehicle as the downtime and maintenance of the fleet costs a lot for fleet companies. Recently with the clean air act, the government is promoting the use of alternative fuel vehicles as they provide soot-free emissions. However, the lack of understanding of how the maintenance cost associated with alternative fuel vehicles changes over time is making fleet companies opt for

diesel vehicles as diesel vehicles are known to be robust for a long time. The studies related to maintenance costs for alternative fuel vehicles have been a challenge due to the lack of availability of data. In this study, WVU in collaboration with fleet management companies has collected large volumes of data related to diesel and alternative fuels performing various tasks such as school buses, delivery trucks, vocational trucks, refuse trucks, transit buses, and goods movement.

Regular machine learning models do not generalize well for real-world complex data involving clustered longitudinal data. The maintenance data collected is complex as the activity performed by each truck involves a different duty cycle impacting the maintenance and the performance of vehicles differently. Hence each of the diesel, natural gas, and propane fuel types has different clusters of data among them associated with the truck activity type. Furthermore, different fleet companies maintain vehicles differently further making the pattern more complex.

To address the challenge and fill the knowledge gap, a mixed effect random forest (MERF) model is developed to capture the complex patterns within the group and between groups taking the overall population distribution into account. The model is fitted using the EM algorithm, allowing us to learn fixed effects and random effects. The model is evaluated on the unseen test data from each cluster and observed to perform well giving the predicted values close to actual values for most of the cases. For scenarios where there is large variation within the cluster, the model seems to perform reasonably. The goal of this study is to develop a generalized model that could capture the random effects in data rather than having an individual model for different activity types using different fuel types for which large volumes of data may not be available. Based on the performance metric achieved by the MERF model on the test dataset indicate that the model is generalized well for clusters seen during the training process as well as to the clusters not seen during the training. Given the fuel type, activity, region of operation, etc., the model predicts the average cost per mile as the age and total miles of operation for truck increases helping the fleet management companies to make procurement decisions.

Chapter 7

Contributions & Future work

The operational costs of a truck play a key role in the total cost of ownership (TCO) for a fleet company. The break-even cost per mile can be broken into various factors such as maintenance & repairs, fuel, driver wages, insurance, etc., of which fuel and maintenance & repairs contribute a lot to the total cost of ownership. Since fuel costs and maintenance & repairs are the two largest factors for the total cost of ownership, examining these factors help in identifying the differences in the total cost of operation between fuels based on duty cycles. To make procurement decisions, fleet management companies should have a clear understanding of how the differences in fuel, and maintenance impact the TCO. Also, with the increasing prices of diesel, optimizing fuel consumption by considering the factors that result in higher fuel consumption for existing vehicles would be beneficial. Hence, in this study, the following contributions are made that would benefit fleet management companies to analyze the operational costs for better procurement decisions as well as to the economy through energy saving.

7.1 Contributions

1. Developed a Neural Network model by identifying and using very few vehicle parameters that affect fuel consumption most and can be easily obtained from vehicle trip telemetry data.

2. Developed a Super-Learner model by fusing Machine Learning models and improved the performance and generalization of model.

3. Analyzed Mixed Effects in maintenance data and Developed Mixed Effects Random Forest

model, a single generalized model that can be used to predict cost-per-mile for HDVs performing different activities and fuel types.

4. Created a database by collecting real-time medium- and heavy-duty vehicle telemetry activity data and maintenance data in collaboration with fleet management companies.

5. Modeled maintenance data for predicting maintenance cost of diesel and alternative fuel vehicles performing different activities that enable companies to make procurement decisions.

6. The study would positively impact the alternative fuel infrastructure stakeholders and AFV manufacturers by helping them identify the barriers associated with AFV adoption in certain vocations and regions

7. This work helps in the diffusion of Alternative Fuel Vehicles in the market.

Our findings can be used by analysts, researchers, and policymakers when determining the relative ownership costs of alternative fuel vehicles (AFVs), optimizing the fuel consumption by trips and deciding directions for future research, and by consumers and fleet operators to select cost-effective vehicles. Future Work

7.2 Future Work

Though alternative fuels are gaining popularity, their usage has not reached the expectation yet. With the current data capturing technologies, it is recommended that companies should concentrate on logging the maintenance data frequently for vehicles using advanced technologies/alternative fuels for longer duration or mileage. This would help in understanding the performance and maintenance of vehicles with different duty cycles using different fuel engines which further helps promote the usage of alternative fuels where feasible to offer soot-free emissions and environmentally reduce pollution.

Including more features such as the maximum and minimum temperatures, precipitation, road grade, etc would enable the development of more robust prediction models. The developed models can be applied to electric vehicles. Models for identifying the factors resulting in high maintenance costs could be beneficial for fleet management companies in understanding the impact of different

parameters on the maintenance of vehicles. Metrics such as Jaccard Similarity or Euclidean Distance are used to identify similarity in data to determine if the model can be used on new clusters of data, for example, new fuel type.

Appendix A

Related Publications and Presentations

A.1 Published

- S. Katreddi and A. Thiruvengadam, "Trip-Based Modeling of Fuel Consumption in Heavy-Duty Vehicles using Artificial Intelligence," *Energies*, Dec. 2021.
- S. Katreddi, S. Kasani and A. Thiruvengadam, "A Review on Applications of Artificial Intelligence in Heavy Duty Trucks," *Energies*, Sep. 2022.

A.2 In Review

- S. Katreddi, A. Thiruvengadam, G. Thompson, N. Schmid, and V. Padmanaban, "Machine Learning Models for Maintenance Cost Estimation in Delivery Trucks using Diesel and Natural Gas Fuels" *Frontiers in Mechanical Engineering.*
- S. Katreddi, A. Thiruvengadam, G. Thompson and N. Schmid, "Mixed Effects Random Forest Model for predicting maintenance cost in heavy-duty vocational vehicles using diesel and alternative fuels" *IEEE Access.*
- S. Katreddi, G. Doretto and D. Adjeroh, "Plant Stress Classification using color channel combination" *Scientific Reports.*

A.3 Conference Presentations

- "Fuel Consumption Estimation in Heavy Duty Vehicles Using Machine Learning." *8th International Conference on Mechanical, Materials, and Manufacturing*, 2021.
- "Semiconductor defect wafer map classification using CNN for yield improvement in semiconductor fabrication" *3rd Asia Digital Image Processing Conference (ADIP 2021)*, 2021.
- "Classification and Analysis of Nanostructure SEM Images using Deep Learning for Material Science Applications" *The 4th International Conference on Control and Computer vision.*

References

[1] O. US EPA, "Draft inventory of u.s. greenhouse gas emissions and sinks: 1990-2021," 2 2023. [Online; accessed 2023-02-28].

[2] "Reduce climate change." [Online; accessed 2023-02-28].

[3] A. Tiwari, "Difference between ai vs ml vs dl," 1 2021. [Online; accessed 2022-09-05].

[4] S. Raschka, *Python Machine Learning*. Birmingham, UK: Packt Publishing, 1st ed., 2015.

[5] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Computer Science*, vol. 2, p. 160, 3 2021.

[6] A. Tch, "The mostly complete chart of neural networks, explained," 2020.

[7] "Application of ai - javatpoint." [Online; accessed 2022-09-05].

[8] A. Ghosh, M. Nashaat, J. Miller, S. Quader, and C. Marston, "A comprehensive review of tools for exploratory analysis of tabular industrial datasets," *Visual Informatics*, vol. 2, no. 4, pp. 235–253, 2018.

[9] N. Wei, Q. Zhang, Y. Zhang, J. Jin, J. Chang, Z. Yang, C. Ma, Z. Jia, C. Ren, L. Wu, J. Peng, and H. Mao, "Super-learner model realizes the transient prediction of co2 and nox of diesel trucks: Model development, evaluation and interpretation," *Environment International*, vol. 158, p. 106977, 1 2022.

[10] Y. Chen and A. Meier, "Fuel consumption impacts of auto roof racks," *Energy Policy*, vol. 92, no. C, pp. 325–333, 2016. publisher: Elsevier.

[11] C. Bai, P. Dallasega, G. Orzes, and J. Sarkis, "Industry 4.0 technologies assessment: A sustainability perspective," *International Journal of Production Economics*, vol. 229, p. 107776, 11 2020.

[12] K. Bheemaiah, M. Esposito, and T. Tse, "What is machine learning?." [Online; accessed 2022-09-05].

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 5 2015. number: 7553 publisher: Nature Publishing Group.

[14] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Handwritten digit recognition with a back-propagation network," vol. 2, Morgan-Kaufmann, 1989. [Online; accessed 2022-05-14].

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," vol. 25, Curran Associates, Inc., 2012. [Online; accessed 2022-05-14].

[16] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," (Paris, France), pp. 253–256, IEEE, 5 2010. [Online; accessed 2022-07-09].

[17] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 10 1986. number: 6088 publisher: Nature Publishing Group.

[18] A. K. Tyagi and S. U. Aswathy, "Autonomous intelligent vehicles (aiv): Research statements, open issues, challenges and road for future," *International Journal of Intelligent Networks*, vol. 2, pp. 83–102, 1 2021.

[19] A. Areiqat, A. Hamdan, A. Alheet, and B. Alareeni, "Impact of artificial intelligence on e-commerce development," pp. 571–578, 3 2021. DOI: 10.1007/978-3-030-69221-6_43.

[20] S. Secinaro, D. Calandra, A. Secinaro, V. Muthurangu, and P. Biancone, "The role of artificial intelligence in healthcare: a structured literature review," *BMC Medical Informatics and Decision Making*, vol. 21, p. 125, 4 2021.

[21] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: an introduction," *Journal of the American Medical Informatics Association*, vol. 18, pp. 544–551, 9 2011.

[22] S. Satpathy, "Role of artificial intelligence in social media marketing," *International Journal of Recent Technology and Engineering*, 1 2019. [Online; accessed 2022-09-05].

[23] R. Benabdelouahed and C. Dakouan, "The use of artificial intelligence in social media: Opportunities and perspectives," p. 6.

[24] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?," 7 2022. arXiv:2207.08815 [cs, stat].

[25] O. US EPA, "Clean trucks plan," 8 2021. [Online; accessed 2022-09-29].

[26] S. C. Davis, S. W. Diegel, and R. G. Boundy, "Transportation energy data book: Edition 30," tech. rep., 7 2011. DOI: 10.2172/1020812.

[27] G. Sullivan, R. Pugh, A. P. Melendez, and W. D. Hunt, "Operations and maintenance best practices - a guide to achieving operational efficiency (release 3)," tech. rep., 8 2010. DOI: 10.2172/1034595.

[28] A. Ali, Q. Si, B. Wang, J. Yuan, P. Wang, G. Rasool, A. Shokrian, A. Ali, and M. A. Zaman, "Comparison of empirical models using experimental results of electrical submersible pump under two-phase flow: numerical and empirical model validation," *Physica Scripta*, vol. 97, p. 065209, 5 2022. publisher: IOP Publishing.

[29] J. Ziółkowski, M. Oszczypała, J. Szkutnik-Rogoż, and J. Malachowski, "Use of artificial neural networks to predict fuel consumption on the basis of technical parameters of vehicles," *Energies*, vol. 14, pp. 1–23, 5 2021.

[30] H. Heni, S. Arona Diop, J. Renaud, and L. C. Coelho, "Measuring fuel consumption in vehicle routing: new estimation models using supervised learning," *International Journal of Production Research*, vol. 0, pp. 1–17, 7 2021. publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00207543.2021.1948133.

[31] S. Zargarnezhad, R. Dashti, and R. Ahmadi, "Predicting vehicle fuel consumption in energy distribution companies using anns," *Transportation Research Part D: Transport and Environment*, vol. 74, pp. 174–188, 9 2019.

[32] E. Moradi and L. Miranda-Moreno, "Vehicular fuel consumption estimation using real-world measures through cascaded machine learning modeling," *Transportation Research Part D: Transport and Environment*, vol. 88, p. 102576, 11 2020.

[33] Y. Du, J. Wu, S. Yang, and L. Zhou, "Predicting vehicle fuel consumption patterns using floating vehicle data," *Journal of Environmental Sciences*, vol. 59, pp. 24–29, 9 2017.

[34] A. Parlak, Y. Islamoglu, H. Yasar, and A. Egrisogut, "Application of artificial neural network to predict specific fuel consumption and exhaust temperature for a diesel engine," *Applied Thermal Engineering*, vol. 26, pp. 824–828, 6 2006.

[35] Y. Yao, X. Zhao, C. Liu, J. Rong, Y. Zhang, Z. Dong, and Y. Su, "Vehicle fuel consumption prediction method based on driving behavior data collected from smartphones," *Journal of Advanced Transportation*, vol. 2020, p. e9263605, 3 2020. publisher: Hindawi.

[36] F. Perrotta, T. Parry, and L. Neves, "Application of machine learning for fuel consumption modelling of trucks," *2017 IEEE International Conference on Big Data (Big Data)*, 2017.

[37] S. Katreddi and A. Thiruvengadam, "Trip based modeling of fuel consumption in modern heavy-duty vehicles using artificial intelligence," *Energies*, vol. 14, p. 8592, 1 2021. number: 24 publisher: Multidisciplinary Digital Publishing Institute.

[38] E. Siami-Irdemoosa and S. R. Dindarloo, "Prediction of fuel consumption of mining dump trucks: A neural networks approach," *Applied Energy*, vol. 151, no. C, pp. 77–84, 2015. publisher: Elsevier.

[39] A. Soofastaei, S. Aminossadati, M. Kizil, and P. Knights, "Reducing fuel consumption of haul trucks in surface mines using artificial intelligence models," 2 2016.

[40] V. Bodell, L. Ekstrom, and S. Aghanavesi, "Comparing machine learning estimation of fuel consumption of heavy-duty vehicles," *International Journal of Transport and Vehicle Engineering*, vol. 15, pp. 97–101, 1 2021.

[41] O. Wysocki, L. Deka, D. Elizondo, J. Kropiwnicki, and J. Czyzewicz, "Heavy duty vehicle fuel consumption modelling based on exploitation data by using artificial neural networks," 2019.

[42] A. Schoen, A. Byerly, B. Hendrix, R. M. Bagwe, d. E. C. Santos, and Z. B. Miled, "A machine learning model for average fuel consumption in heavy vehicles," *IEEE Transactions on Vehicular Technology*, vol. 68, pp. 6343–6351, 7 2019. event-title: IEEE Transactions on Vehicular Technology.

[43] A. Barbado and O. Corcho, "Vehicle fuel optimization under real-world driving conditions: An explainable artificial intelligence approach.," p. 30.

[44] R. Zhang, Y. Wang, Y. Pang, B. Zhang, Y. Wei, M. Wang, and R. Zhu, "A deep learning micro-scale model to estimate the co2 emissions from light-duty diesel trucks based on real-world driving," *Atmosphere*, vol. 13, p. 1466, 9 2022. number: 9 publisher: Multidisciplinary Digital Publishing Institute.

[45] C. Le Cornec, N. Molden, M. Van Reeuwijk, and M. Stettler, "Modelling of instantaneous emissions from diesel vehicles with a special focus on nox: Insights from machine learning techniques," *Science of The Total Environment*, vol. 737, p. 139625, 5 2020.

[46] N. Palanichamy, S.-C. Haw, S. S, R. Murugan, and K. Govindasamy, "*Machine learning methods to predict particulate matter $PM_{2.5}$*," tech. rep., 4 2022. DOI: 10.12688/f1000research.73166.1 type: article.

[47] M. Danesh Yazdi, Z. Kuang, K. Dimakopoulou, B. Barratt, E. Suel, H. Amini, A. Lyapustin, K. Katsouyanni, and J. Schwartz, "Predicting fine particulate matter (pm2.5) in the greater london area: An ensemble approach using machine learning methods," *Remote Sensing*, vol. 12, p. 914, 1 2020. number: 6 publisher: Multidisciplinary Digital Publishing Institute.

[48] H.-T. Wen, J.-H. Lu, and D.-S. Jhang, "Features importance analysis of diesel vehicles' nox and co2 emission predictions in real road driving based on gradient boosting regression model," *International Journal of Environmental Research and Public Health*, vol. 18, p. 13044, 12 2021. PMID: 34948649 PMCID: PMC8700826.

[49] S. Bhowmik, A. Paul, R. Panua, S. K. Ghosh, and D. Debroy, "Performance-exhaust emission prediction of diesosenol fueled diesel engine: An ann coupled morsm based optimization," *Energy*, vol. 153, pp. 212–222, 6 2018.

[50] S. Roy, R. Banerjee, and P. K. Bose, "Performance and exhaust emissions prediction of a crdi assisted single cylinder diesel engine coupled with egr using artificial neural network," *Applied Energy*, vol. 119, pp. 330–340, 4 2014.

[51] B. R. Hosamani, S. Abbas Ali, and V. Katti, "Assessment of performance and exhaust emission quality of different compression ratio engine using two biodiesel mixture: Artificial neural network approach," *Alexandria Engineering Journal*, vol. 60, pp. 837–844, 2 2021.

[52] C. Maino, D. Misul, A. Di Mauro, and E. Spessa, "A deep neural network based model for the prediction of hybrid electric vehicles carbon dioxide emissions," *Energy and AI*, vol. 5, p. 100073, 9 2021.

[53] J. Seo, B. Yun, J. Park, J. Park, M. Shin, and S. Park, "Prediction of instantaneous real-world emissions from diesel light-duty vehicles based on an integrated artificial neural network and vehicle dynamics model," *The Science of the total environment*, vol. 786, p. 147359, 9 2021. PMID: 33964768.

[54] C. Tóth-Nagy, J. J. Conley, R. P. Jarrett, and N. N. Clark, "Further validation of artificial neural network-based emissions simulation models for conventional and hybrid electric vehicles," *Journal of the Air & Waste Management Association*, vol. 56, pp. 898–910, 7 2006.

[55] O. Azeez, B. Pradhan, H. Shafri, N. Shukla, and H. Rizeei, "Modeling of co emissions from traffic vehicles using artificial neural networks," *Applied Sciences*, vol. 9, p. 313, 1 2019.

[56] S. Khurana, S. Saxena, S. Jain, and A. Dixit, "Predictive modeling of engine emissions using machine learning: A review," *Materials Today: Proceedings*, vol. 38, pp. 280–284, 1 2021.

[57] R. Pillai, V. Triantopoulos, A. S. Berahas, M. Brusstar, R. Sun, T. Nevius, and A. L. Boehman, "Modeling and predicting heavy-duty vehicle engine-out and tailpipe nitrogen oxide (nox) emissions using deep learning," *Frontiers in Mechanical Engineering*, vol. 8, 2022. [Online; accessed 2022-05-24].

[58] J. Mohammadhassani, S. Khalilarya, M. Solimanpur, and A. Dadvand, "Prediction of nox emissions from a direct injection diesel engine using artificial neural network," *Modelling and Simulation in Engineering*, vol. 2012, p. e830365, 6 2012. publisher: Hindawi.

[59] v. d. M. J. Laan, E. C. Polley, and A. E. Hubbard, "Super learner," *Statistical Applications in Genetics and Molecular Biology*, vol. 6, 9 2007. publisher: De Gruyter.

[60] Y. Yu, Y. Wang, J. Li, M. Fu, A. N. Shah, and C. He, "A novel deep learning approach to predict the instantaneous no emissions from diesel engine," *IEEE Access*, vol. 9, pp. 11002–11013, 2021.

[61] J. Gao, J. Yi, H. Zhu, and Y. L. Murphey, "A personalized lane-changing model for advanced driver assistance system based on deep learning and spatial-temporal modeling," *SAE International Journal of Transportation Safety*, vol. 7, no. 2, pp. 163–174, 2019. publisher: SAE International.

[62] Q. Wang, W. Zhuang, L. Wang, and F. Ju, "Lane keeping assist for an autonomous vehicle based on deep reinforcement learning," tech. rep., Warrendale, PA, 4 2020. ISSN: 0148-7191, 2688-3627 DOI: 10.4271/2020-01-0728.

[63] Z. Wei, C. Wang, P. Hao, and M. J. Barth, "Vision-based lane-changing behavior detection using deep residual neural network," (Auckland, New Zealand), pp. 3108–3113, IEEE, 10 2019. [Online; accessed 2022-06-21].

[64] V. Mahajan, C. Katrakazas, and C. Antoniou, "Prediction of lane-changing maneuvers with automatic labeling and deep learning," *Transportation Research Record*, vol. 2674, pp. 336–347, 7 2020. publisher: SAGE Publications Inc.

[65] M. Karthikeyan, S. Sathiamoorthy, and M. Vasudevan, "Lane keep assist system for an autonomous vehicle using support vector machine learning algorithm," Lecture Notes on Data Engineering and Communications Technologies, (Cham), pp. 101–108, Springer International Publishing, 2020.

[66] L. Zhao, "Stereo- and neural network-based pedestrian detection," *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS*, vol. 1, no. 3, p. 7, 2000.

[67] P. J. Navarro, C. Fernández, R. Borraz, and D. Alonso, "A machine learning approach to pedestrian detection for autonomous vehicles using high-definition 3d range data," *Sensors (Basel, Switzerland)*, vol. 17, p. 18, 12 2016. PMID: 28025565 PMCID: PMC5298591.

[68] M. M. Islam, A. A. R. Newaz, and A. Karimoddini, "A pedestrian detection and tracking framework for autonomous cars: Efficient fusion of camera and lidar data," (Melbourne, Australia), pp. 1287–1292, IEEE, 10 2021. [Online; accessed 2022-06-21].

[69] J. Cao, C. Song, S. Peng, S. Song, X. Zhang, Y. Shao, and F. Xiao, "Pedestrian detection algorithm for intelligent vehicles in complex scenarios," *Sensors*, vol. 20, p. 3646, 1 2020. number: 13 publisher: Multidisciplinary Digital Publishing Institute.

[70] A. Angelova, A. Krizhevsky, and V. Vanhoucke, "Pedestrian detection with a large-field-of-view deep network," (Seattle, WA, USA), pp. 704–711, IEEE, 5 2015. [Online; accessed 2022-06-21].

[71] V. Ortiz Castelló, O. del Tejo Catalá, I. Salvador Igual, and J.-C. Perez-Cortes, "Real-time on-board pedestrian detection using generic single-stage algorithms and on-road databases," *International Journal of Advanced Robotic Systems*, vol. 17, p. 1729881420929175, 9 2020. publisher: SAGE Publications.

[72] H. Song, H. Liang, H. Li, Z. Dai, and X. Yun, "Vision-based vehicle detection and counting system using deep learning in highway scenes," *European Transport Research Review*, vol. 11, p. 51, 12 2019.

[73] H. Herunde, A. Singh, H. Deshpande, and P. Shetty, "Detection of pedestrian and different types of vehicles using image processing," *International Journal of Research in Industrial Engineering*, 5 2020. [Online; accessed 2022-06-21].

[74] L. G. Galvao, M. Abbod, T. Kalganova, V. Palade, and M. N. Huda, "Pedestrian and vehicle detection in autonomous vehicle perception systems—a review," *Sensors*, vol. 21, p. 7267, 10 2021.

[75] A. Gupta, A. Anpalagan, L. Guan, and A. S. Khwaja, "Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues," *Array*, vol. 10, p. 100057, 7 2021.

[76] Z. Li, Q. Zeng, Y. Liu, J. Liu, and L. Li, "An improved traffic lights recognition algorithm for autonomous driving in complex scenarios," *International Journal of Distributed Sensor Networks*, vol. 17, p. 15501477211018374, 5 2021. publisher: SAGE Publications.

[77] G. Mu, Z. Xinyu, L. Deyi, Z. Tianlei, and A. Lifeng, "Traffic light detection and recognition for autonomous vehicles," *The Journal of China Universities of Posts and Telecommunications*, vol. 22, 2 2015.

[78] S. Swetha and P. Sivakumar, "Ssla based traffic sign and lane detection for autonomous cars," pp. 766–771, 3 2021.

[79] M. Cunneen, M. Mullins, and F. Murphy, "Autonomous vehicles and embedded artificial intelligence: The challenges of framing machine driving decisions," *Applied Artificial Intelligence*, vol. 33, pp. 706–731, 7 2019. publisher: Taylor & Francis _eprint: https://doi.org/10.1080/08839514.2019.1600301.

[80] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, "Explainable artificial intelligence for autonomous driving: An overview and guide for future research directions," 4 2022. number: arXiv:2112.11561 arXiv:2112.11561 [cs].

[81] G. Lugano, "Virtual assistants and self-driving cars," pp. 1–5, 5 2017.

[82] J. Jagelčák, J. Gnap, O. Kuba, J. Frnda, and M. Kostrzewski, "Determination of turning radius and lateral acceleration of vehicle by gnss/ins sensor," *Sensors (Basel, Switzerland)*, vol. 22, p. 2298, 3 2022. PMID: 35336468 PMCID: PMC8950859.

[83] Z. Zhou, Z. Akhtar, K. L. Man, and K. Siddique, "A deep learning platooning-based video information-sharing internet of things framework for autonomous driving systems," *International Journal of Distributed Sensor Networks*, vol. 15, p. 1550147719883133, 11 2019. publisher: SAGE Publications.

[84] "This year, autonomous trucks will take to the road with no one on board," 1 2021. section: Transportation.

[85] M. Song, F. Chen, and X. Ma, "Organization of autonomous truck platoon considering energy saving and pavement fatigue," *Transportation Research Part D: Transport and Environment*, vol. 90, p. 102667, 1 2021.

[86] S. Tsugawa, S. Jeschke, and S. E. Shladover, "A review of truck platooning projects for energy savings," *IEEE Transactions on Intelligent Vehicles*, vol. 1, pp. 68–77, 3 2016. event-title: IEEE Transactions on Intelligent Vehicles.

[87] S. Tsugawa, S. Kato, and K. Aoki, "An automated truck platoon for energy saving," pp. 4109–4114, 9 2011. ISSN: 2153-0866.

[88] F. Jaffar, T. Farid, M. Sajid, Y. Ayaz, and M. J. Khan, "Prediction of drag force on vehicles in a platoon configuration using machine learning," *IEEE Access*, vol. 8, pp. 201823–201834, 2020. event-title: IEEE Access.

[89] J. Yang, W. Peng, and C. Sun, "A learning control method of automated vehicle platoon at straight path with ddpg-based pid," *Electronics*, vol. 10, p. 2580, 1 2021. number: 21 publisher: Multidisciplinary Digital Publishing Institute.

[90] C. Zhang, Z. Ouyang, L. Ren, and Y. Liu, "Low-cost lidar-based vehicle detection for self-driving container trucks at seaport," Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, (Cham), pp. 451–466, Springer International Publishing, 2021.

[91] D. Wang, L. Gao, Z. Lan, W. Li, J. Ren, J. Zhang, P. Zhang, P. Zhou, S. Wang, J. Pan, D. Manocha, and R. Yang, "An intelligent self-driving truck system for highway transportation," *Frontiers in Neurorobotics*, vol. 16, p. 843026, 5 2022.

[92] Y. Ran, X. Zhou, P. Lin, Y. Wen, and R. Deng, "A survey of predictive maintenance: Systems, purposes and approaches," tech. rep., 12 2019. arXiv:1912.07383 [cs, eess] type: article.

[93] T. P. Carvalho, F. A. A. M. N. Soares, R. Vita, R. d. P. Francisco, J. P. Basto, and S. G. S. Alcalá, "A systematic literature review of machine learning methods applied to predictive maintenance," *Computers & Industrial Engineering*, vol. 137, p. 106024, 11 2019.

[94] A. Theissler, J. Pérez-Velázquez, M. Kettelgerdes, and G. Elger, "Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry," *Reliability Engineering & System Safety*, vol. 215, p. 107864, 11 2021.

[95] A. Chaudhuri, "Predictive maintenance for industrial iot of vehicle fleets using hierarchical modified fuzzy support vector machine," p. 15.

[96] F. Arena, M. Collotta, L. Luca, M. Ruggieri, and F. Termine, "Predictive maintenance in the automotive sector: A literature review," *Mathematical and Computational Applications*, 2021.

[97] R. Prytz, S. Nowaczyk, T. Rögnvaldsson, and S. Byttner, "Predicting the need for vehicle compressor repairs using maintenance records and logged vehicle data," *Engineering Applications of Artificial Intelligence*, vol. 41, pp. 139–150, 5 2015.

[98] V. Revanur, A. Ayibiowu, M. Rahat, and R. Khoshkangini, "Embeddings based parallel stacked autoencoder approach for dimensionality reduction and predictive maintenance of vehicles," Communications in Computer and Information Science, (Cham), pp. 127–141, Springer International Publishing, 2020.

[99] C. Chen, Y. Liu, X. Sun, C. D. Cairano-Gilfedder, and S. Titmus, "Automobile maintenance prediction using deep learning with gis data," *Procedia CIRP*, vol. 81, pp. 447–452, 2019.

[100] Y. Sun, Z. Xu, and T. Zhang, "On-board predictive maintenance with machine learning," tech. rep., Warrendale, PA, 4 2019. ISSN: 0148-7191, 2688-3627 DOI: 10.4271/2019-01-1048.

[101] D. Rengasamy, M. Jafari, B. Rothwell, X. Chen, and G. P. Figueredo, "Deep learning with dynamically weighted loss function for sensor-based prognostics and health management," *Sensors*, vol. 20, p. 723, 1 2020. number: 3 publisher: Multidisciplinary Digital Publishing Institute.

[102] J. Wang and H. A. Rakha, "Fuel consumption model for conventional diesel buses," *Applied Energy*, vol. 170, pp. 394–402, 5 2016.

[103] D. Zhu and X. Zheng, "Fuel consumption and emission characteristics in asymmetric twin-scroll turbocharged diesel engine with two exhaust gas recirculation circuits," *Applied Energy*, vol. 238, pp. 985–995, 3 2019.

[104] C.-L. Lo, C.-H. Chen, T.-S. Kuan, K.-R. Lo, and H.-J. Cho, "Fuel consumption estimation system and method with lower cost," *Symmetry*, vol. 9, p. 105, 7 2017. number: 7 publisher: Multidisciplinary Digital Publishing Institute.

[105] E. Hellström, M. Ivarsson, J. Åslund, and L. Nielsen, "Look-ahead control for heavy trucks to minimize trip time and fuel consumption," *Control Engineering Practice*, vol. 17, pp. 245–254, 2 2009.

[106] O. US EPA, "Sources of greenhouse gas emissions," 12 2015. [Online; accessed 2022-05-23].

[107] S. T. Anderson, I. W. H. Parry, J. M. Sallee, and C. Fischer, "Automobile fuel economy standards: Impacts, efficiency, and alternatives," p. 20.

[108] "Corporate average fuel economy | nhtsa." [Online; accessed 2022-05-23].

[109] S. H. Nasser, V. Weibermel, and J. Wiek, "Computer simulation of vehicle's performance and fuel consumption under steady and dynamic driving conditions," tech. rep., Warrendale, PA, 2 1998. ISSN: 0148-7191, 2688-3627 DOI: 10.4271/981089.

[110] A. Cappiello, I. Chabini, E. Nam, A. Lue, and M. Abou Zeid, "A statistical model of vehicle emissions and fuel consumption," (Singapore), pp. 801–809, IEEE, 2002. [Online; accessed 2022-09-06].

[111] I. Ivkovic, S. Kaplanović, and B. Milovanović, "Influence of road and traffic conditions on fuel consumption and fuel cost for different bus technologies," *Thermal Science*, vol. 2017, pp. 693–706, 1 2017.

[112] G. N. Bifulco, F. Galante, L. Pariota, and M. R. Spena, "A linear model for the estimation of fuel consumption and the impact evaluation of advanced driving assistance systems," *Sustainability*, vol. 7, pp. 14326–14343, 10 2015. number: 10 publisher: Multidisciplinary Digital Publishing Institute.

[113] Y. Yao, X. Zhao, Y. Zhang, C. Chen, and J. Rong, "Modeling of individual vehicle safety and fuel consumption under comprehensive external conditions," *Transportation Research Part D: Transport and Environment*, vol. 79, p. 102224, 2 2020.

[114] F. Zannikos, E. Tzirakis, and S. Stournas, *Impact of driving style on fuel consumption and exhaust emissions: defensive and aggressive driving style*. 9 2007.

[115] H. Ma, H. Xie, D. Huang, and S. Xiong, "Effects of driving style on the fuel consumption of city buses under different road conditions and vehicle masses," *Transportation Research Part D: Transport and Environment*, vol. 41, pp. 205–216, 12 2015.

[116] R. Shang, Y. Zhang, Z.-J. Max Shen, and Y. Zhang, "Analyzing the effects of road type and rainy weather on fuel consumption and emissions: A mesoscopic model based on big traffic data," *IEEE Access*, vol. 9, pp. 62298–62315, 2021.

[117] H. Jeon, "The impact of climate change on passenger vehicle fuel consumption: Evidence from u.s. panel data," *Energies*, vol. 12, no. 23, 2019.

[118] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 9 1995.

[119] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 10 2001.

[120] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, pp. 386–408, 11 1958. PMID: 13602029.

[121] W. S. Mcculloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," p. 17.

[122] Y.-R. Kim, M. Jung, and J.-B. Park, "Development of a fuel consumption prediction model based on machine learning using ship in-service data," *Journal of Marine Science and Engineering*, vol. 9, p. 137, 2 2021. number: 2 publisher: Multidisciplinary Digital Publishing Institute.

[123] S. Baumann, T. Neidhardt, and U. Klingauf, "Evaluation of the aircraft fuel economy using advanced statistics and machine learning," *CEAS Aeronautical Journal*, vol. 12, pp. 669–681, 8 2021.

[124] A. Amer, A. Abdalla, A. Noraziah, and A. A. C. Fauzi, "Prediction of vehicle fuel consumption model based on artificial neural network," *Applied Mechanics and Materials*, vol. 492, pp. 3–6, 2014. event-title: Power and Energy Systems III ISBN: 9783037859919 publisher: Trans Tech Publications Ltd.

[125] D. Freedman, *Statistical Models: Theory and Practice*. 1 2005. journalAbbreviation: Statistical Models: Theory and Practice container-title: Statistical Models: Theory and Practice DOI: 10.1017/CBO9781139165495.

[126] M. J. Barth, F. An, T. Younglove, G. Scora, C. Levine, M. Ross, and T. Wenzel, "Development of a comprehensive modal emissions model," *NCHRP Web-Only Document*, 4 2000. number: NCHRP Project 25-11.

[127] C. John, R. Friedrich, J. Staehelin, K. Schläpfer, and W. A. Stahel, "Comparison of emission factors for road traffic from a tunnel study (gubrist tunnel, switzerland) and from emission modeling," *Atmospheric Environment*, vol. 33, pp. 3367–3376, 9 1999.

[128] G. Song, L. Yu, and X. Zhang, "Emission analysis at toll station area in beijing with portable emission measurement system," *Transportation Research Record*, vol. 2058, pp. 106–114, 1 2008. publisher: SAGE Publications Inc.

[129] H. Liu, M. Barth, G. Scora, N. Davis, and J. Lents, "Using portable emission measurement systems for transportation emissions studies: Comparison with laboratory methods," *Transportation Research Record*, vol. 2158, pp. 54–60, 1 2010. publisher: SAGE Publications Inc.

[130] Y. S. Park and S. Lek, "Chapter 7 - artificial neural networks: Multilayer perceptron for ecological modeling," in *Developments in Environmental Modelling* (S. E. Jørgensen, ed.), vol. 28 of *Ecological Model Types*, pp. 123–140, Elsevier, 1 2016. DOI: 10.1016/B978-0-444-63623-2.00007-4.

[131] R. Lippmann, "An introduction to computing with neural nets," *IEEE ASSP Magazine*, vol. 4, pp. 4–22, 4 1987. event-title: IEEE ASSP Magazine.

[132] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, pp. 359–366, 1 1989.

[133] J. Brownlee, "A gentle introduction to the rectified linear unit (relu)," 1 2019. [Online; accessed 2022-09-06].

[134] Y. Zhang and Q. Liu, "Reliability-based design of automobile components," *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, vol. 216, pp. 455–471, 6 2002.

[135] O. US EPA, "Fast facts on transportation greenhouse gas emissions," 8 2015. [Online; accessed 2022-09-18].

[136] M. A. Ghadikolaei, P. K. Wong, C. S. Cheung, J. Zhao, Z. Ning, K.-F. Yung, H. C. Wong, and N. K. Gali, "Why is the world not yet ready to use alternative fuel vehicles?," *Heliyon*, vol. 7, p. e07527, 7 2021.

[137] O. US EPA, "Learn about impacts of diesel exhaust and the diesel emissions reduction act (dera)," 7 2015. [Online; accessed 2023-02-16].

[138] C. Jang and J. Lee, "Experimental investigation of the effects of various factors on the emission characteristics of low-emission natural gas vehicles," *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, vol. 219, pp. 825–831, 6 2005. publisher: IMECHE.

[139] E. R. Jayaratne, Z. D. Ristovski, N. Meyer, and L. Morawska, "Particle and gaseous emissions from compressed natural gas and ultralow sulphur diesel-fuelled buses at four steady engine loads," *Science of The Total Environment*, vol. 407, pp. 2845–2852, 4 2009.

[140] M. Hearst, S. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, vol. 13, pp. 18–28, 7 1998. event: IEEE Intelligent Systems and their Applications.

[141] G. Bebis and M. Georgiopoulos, "Feed-forward neural networks," *IEEE Potentials*, vol. 13, pp. 27–31, 10 1994. event: IEEE Potentials.

[142] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. Adaptive computation and machine learning, Cambridge, Mass: MIT Press, 2006. OCLC: ocm61285753.

[143] I. Tessaro, V. C. Mariani, and L. d. S. Coelho, "Machine learning models applied to predictive maintenance in automotive engine components," *Proceedings*, vol. 64, no. 1, p. 26, 2020. number: 1 publisher: Multidisciplinary Digital Publishing Institute.

[144] W. Yu, I. Y. Kim, and C. Mechefske, "An improved similarity-based prognostic algorithm for rul estimation using an rnn autoencoder scheme," *Reliability Engineering & System Safety*, vol. 199, p. 106926, 7 2020.

[145] J. Zhu, N. Chen, and W. Peng, "Estimation of bearing remaining useful life based on multiscale convolutional neural network," *IEEE Transactions on Industrial Electronics*, vol. 66, pp. 3208–3216, 4 2019. event: IEEE Transactions on Industrial Electronics.

[146] A. Malhi, R. Yan, and R. X. Gao, "Prognosis of defect propagation based on recurrent neural networks," *IEEE Transactions on Instrumentation and Measurement*, vol. 60, pp. 703–711, 3 2011. event: IEEE Transactions on Instrumentation and Measurement.

[147] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *The Annals of Applied Statistics*, vol. 2, 9 2008. arXiv: 0811.1645.

[148] S. Voronov, M. Krysander, and E. Frisk, "Predictive maintenance of lead-acid batteries with sparse vehicle operational data," *International Journal of Prognostics and Health Management*, vol. 11, no. 1, 2020. number: 1.

[149] S. Safavi, M. A. Safavi, H. Hamid, and S. Fallah, "Multi-sensor fault detection, identification, isolation and health forecasting for autonomous vehicles," *Sensors*, vol. 21, p. 2547, 1 2021. number: 7 publisher: Multidisciplinary Digital Publishing Institute.

[150] H. Hu, H. Luo, and X. Deng, "Health monitoring of automotive suspensions: A lstm network approach," *Shock and Vibration*, vol. 2021, p. e6626024, 4 2021. publisher: Hindawi.

[151] K. Andrzejczak, M. Młyńczak, and J. Selech, "Poisson-distributed failures in the predicting of the cost of corrective maintenance," *Eksploatacja i Niezawodność*, vol. Vol. 20, no. no. 4, 2018. [Online; accessed 2022-10-12].

[152] A. I. Adekitan, A. Bukola, and O. Kennedy, "A data-based investigation of vehicle maintenance cost components using ann," *IOP Conference Series: Materials Science and Engineering*, vol. 413, p. 012009, 9 2018. publisher: IOP Publishing.

[153] G. Barnes and P. Langworthy, "Per mile costs of operating automobiles and trucks," *Transportation Research Record*, vol. 1864, pp. 71–77, 1 2004. publisher: SAGE Publications Inc.

[154] Y. Chen, J. Ma, B. Han, P. Zhang, H. Hua, H. Chen, and X. Su, "Emissions of automobiles fueled with alternative fuels based on engine technology: A review," *Journal of Traffic and Transportation Engineering (English Edition)*, vol. 5, pp. 318–334, 8 2018.

[155] J. Brownlee, "Why one-hot encode data in machine learning?," 7 2017. [Online; accessed 2023-03-22].

[156] United States. Department of Transportation. Bureau of Transportation Statistics, "National transportation statistics (NTS)," 2019. [Online; accessed 2023-03-22].

[157] Y. Bae, S. K. Mitra, C. R. Rindt, and S. G. Ritchie, "Factors influencing alternative fuel adoption decisions in heavy-duty vehicle fleets," *Transportation Research Part D: Transport and Environment*, vol. 102, p. 103150, 1 2022.

[158] B. Mahesh, *Machine Learning Algorithms -A Review*. 1 2019. DOI: 10.21275/ART20203995.

[159] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, pp. 175–185, 8 1992. publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/00031305.1992.10475879.

[160] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," KDD '16, (New York, NY, USA), p. 785–794, Association for Computing Machinery, 8 2016. [Online; accessed 2022-10-10].

[161] "Alternative fuels data center: Maps and data - vehicle weight classes & categories." [Online; accessed 2023-02-03].

[162] "California moves to accelerate to 100% new zero-emission vehicle sales by 2035 | california air resources board." [Online; accessed 2023-02-18].

[163] "Alternative fuels data center: Alternative fuels and advanced vehicles." [Online; accessed 2023-02-18].

[164] H. Chen, J. He, and X. Zhong, "Engine combustion and emission fuelled with natural gas: A review," *Journal of the Energy Institute*, vol. 92, pp. 1123–1136, 8 2019.

[165] S. Z. Rajper and J. Albrecht, "Prospects of electric vehicles in the developing countries: A literature review," *Sustainability*, vol. 12, p. 1906, 3 2020.

[166] M.-K. Tran, M. Akinsanya, S. Panchal, R. Fraser, and M. Fowler, "Design of a hybrid electric vehicle powertrain for performance optimization considering various powertrain components and configurations," *Vehicles*, vol. 3, no. 1, pp. 20–32, 2021.

[167] M.-K. Tran, S. Sherman, E. Samadani, R. Vrolyk, D. Wong, M. Lowery, and M. Fowler, "Environmental and economic benefits of a battery electric vehicle powertrain with a zinc–air range extender in the transition to electric vehicles," *Vehicles*, vol. 2, no. 3, pp. 398–412, 2020.

[168] "Advanced clean fleets | california air resources board." [Online; accessed 2023-02-18].

[169] L. Mohammed, E. Niesten, and D. Gagliardi, "Adoption of alternative fuel vehicle fleets – a theoretical framework of barriers and enablers," *Transportation Research Part D: Transport and Environment*, vol. 88, p. 102558, 11 2020.

[170] F. Kleiner and H. E. Friedrich, "Maintenance & repair cost calculation and assessment of resale value for different alternative commercial vehicle powertrain technologies," (Stuttgart), 10 2017. [Online; accessed 2022-09-26].

[171] D. Fenton and A. Kailas, "Redefining goods movement: Building an ecosystem for the introduction of heavy-duty battery-electric vehicles," *World Electric Vehicle Journal*, vol. 12, p. 147, 9 2021. number: 3 publisher: Multidisciplinary Digital Publishing Institute.

[172] J. Horrox and M. Casale, "Electric buses in america: Lessons from cities pioneering clean transportation," 10 2019. [Online; accessed 2022-09-26].

[173] N. Williams and D. Murray, "An analysis of the operational costs of trucking: 2020 update," 11 2020. [Online; accessed 2023-02-18].

[174] A. Burnham, D. Gohlke, L. Rush, T. Stephens, Y. Zhou, M. Delucchi, A. Birky, C. Hunter, Z. Lin, S. Ou, F. Xie, C. Proctor, S. Wiryadinata, N. Liu, and M. Boloor, "Comprehensive total cost of ownership quantification for vehicles with different size classes and powertrains," tech. rep., 4 2021. DOI: 10.2172/1780970.

[175] B. Propfe, M. Redelbach, D. J. Santini, and H. Friedrich, "Cost analysis of plug-in hybrid electric vehicles including maintenance & repair costs and resale values," *World Electric Vehicle Journal*, vol. 5, pp. 886–895, 12 2012. number: 4 publisher: Multidisciplinary Digital Publishing Institute.

[176] C. Hunter, M. Penev, E. Reznicek, J. Lustbader, A. Birky, and C. Zhang, "Spatial and temporal analysis of the total cost of ownership for class 8 tractors and class 4 parcel delivery trucks," tech. rep., 9 2021. DOI: 10.2172/1821615.

[177] L. K. Mitropoulos, P. D. Prevedouros, and P. Kopelias, "Total cost of ownership and externalities of conventional, hybrid and electric vehicle," *Transportation Research Procedia*, vol. 24, no. Supplement C, pp. 267 – 274, 2017. 3rd Conference on Sustainable Urban Mobility, 3rd CSUM 2016, 26 â€“ 27 May 2016, Volos, Greece.

[178] J. Hagman, S. Ritzen, J. Janhager, and Y. Susilo, "Total cost of ownership and its potential implications for battery electric vehicle diffusion," *Research in Transportation Business & Management*, vol. 18, 01 2016.

[179] R. Prytz, S. Nowaczyk, T. S. Rögnvaldsson, and S. Byttner, "Analysis of truck compressor failures based on logged vehicle data," in *Industrial Conference on Data Mining*, 2013.

[180] I. C. Boyce, *Comparison of Maintenance Cost of Medium and Heavy-Duty Alternative Fuel and Diesel Vehicles*. PhD thesis, 1 2022. DOI: 10.33915/etd.11586.

[181] M. Borenstein, L. Hedges, J. Higgins, and H. Rothstein, *Introduction to meta-analysis*. wiley, Jan. 2009. Publisher Copyright: © 2009 John Wiley & Sons, Ltd.

[182] L. Grilli and C. Rampichini, "Specification of random effects in multilevel models: a review," *Quality & Quantity: International Journal of Methodology*, vol. 49, no. 3, pp. 967–976, 2015. publisher: Springer.

[183] C. J. M. Maas and J. J. Hox, "Robustness issues in multilevel regression analysis," *Statistica Neerlandica*, vol. 58, pp. 127–137, May 2004.

[184] R. J. Sela and J. S. Simonoff, "Re-em trees: A data mining approach for longitudinal and clustered data," *Mach. Learn.*, vol. 86, p. 169–207, feb 2012.

[185] A. Hajjem, F. Bellavance, and D. Larocque, "Mixed effects regression trees for clustered data," *Statistics & Probability Letters*, vol. 81, no. 4, pp. 451–459, 2011. publisher: Elsevier.

[186] A. D. Gordon, L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and regression trees.," vol. 40, p. 874, 9 1984. ISSN: 0006341X issue: 3 journalAbbreviation: Biometrics.

[187] G. Biau and E. Scornet, "A random forest guided tour," *TEST*, vol. 25, pp. 197–227, 2015.

[188] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123–140, 8 1996.

[189] A. Hajjem, F. Bellavance, and D. Larocque, "Mixed-effects random forest for clustered data," *Journal of Statistical Computation and Simulation*, vol. 84, pp. 1313–1328, 6 2014. publisher: Taylor & Francis eprint: https://doi.org/10.1080/00949655.2012.741599.

[190] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the*EM*algorithm," *J. R. Stat. Soc.*, vol. 39, pp. 1–22, Sept. 1977.

[191] G. J. McLachlan, T. Krishnan, and S. K. Ng, "The em algorithm," Papers 2004,24, 2004.

[192] P. Krennmair and T. Schmid, "Flexible domain prediction using mixed effects random forests," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 71, pp. 1865–1894, 11 2022. arXiv:2201.10933 [stat].