Predictive Maintenance in Automotive Telematics using Machine Learning

Author: Jay Guwalani, ***Senior Data Science Engineer, Bridgestone***
Email: guwalanij3@gmail.com
LinkedIn: linkedin.com/in/jay-guwalani-66763b191

Abstract

Predictive maintenance in automotive telematics signifies a revolutionary method for vehicle health management, using machine learning methods to foresee breakdowns and enhance maintenance schedules. This research utilizes machine learning methods to ascertain the loading status of trucks—loaded or empty—exclusively using data from the vehicle's communication network, particularly from the engine module. We attained an accuracy over 85% for small hauls (0.5 to 5 km) and approximately 95% for long hauls (5 to 500 km). This method optimizes fleet management by minimizing communication between managers and drivers, while also significantly contributing to research on fuel consumption reduction and advanced fault diagnostics. The findings demonstrate that machine learning-based predictive maintenance decreases unplanned downtime and maintenance expenses while also improving vehicle safety and durability. This paper provides a thorough examination of the efficacy of machine learning models in predictive maintenance, delineates the challenges associated with data privacy, computational efficiency, and integration with current automotive systems, and explores future avenues for creating more resilient and scalable predictive maintenance frameworks in the automotive sector.

Keywords: Predictive maintenance, machine learning, vehicle, telematics

1. Introduction

Maintenance is essential since it significantly impacts asset longevity. The operational lifespan of a system may be prolonged by the use of Predictive Maintenance (PdM). Predictive Maintenance (PdM) in the automobile sector demonstrates predictive analytics properly [1]. It assists enterprises in identifying the optimal timing for maintaining a machine or vehicle component by using methods such as data mining, data preparation, and machine learning algorithms [2]. Predictive maintenance employs historical and real-time data from diverse operational components to foresee issues prior to their occurrence [3]. Examples of using predictive maintenance and associated sensors include vibration analysis, oil analysis, thermal imaging, and equipment monitoring. In an automobile production setting, both planned and unexpected downtime may incur significant expenses, leading to substantial setbacks. Predictive maintenance enables continuous real-time monitoring of industrial equipment health and forecasting the likelihood of problems. This enhances operational efficiency and diminishes equipment maintenance costs. The automobile sector is progressively embracing predictive maintenance to prevent unforeseen failures and save maintenance expenses [4]. Figure 1 illustrates the variation in health status between the PdM technique and passive maintenance. Passive maintenance is referred to as run-to-failure. Maintenance is executed when the asset's health state falls to zero. The asset's downtime will be excessively prolonged. In stark contrast, the use of PdM enables the early prediction of failure occurrences and subsequent maintenance implementation, significantly reducing asset downtime [5]. By accurately predicting equipment breakdown times, maintenance may be preemptively planned to reduce the likelihood of accidents, financial losses, and human casualties. Currently, Predictive Maintenance (PdM) is extensively used across several sectors, including automotive [6], aerospace [7], and manufacturing [8]. The use of Predictive Maintenance (PdM) may enhance asset availability by 5% to 15%, decrease maintenance expenditures by 18% to 25%, and minimize machine downtime by 30% to 50% [9]. The car is a kind of costly asset. Automobile fleet management involves overseeing a company's automobiles and vans, which is crucial for logistics.

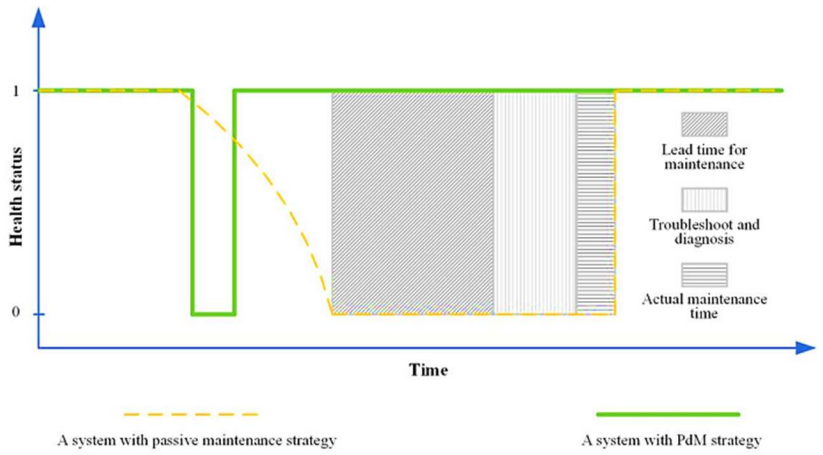

Figure 1. The comparison of passive maintenance and PdM

1.1 Machine learning for Predictive Maintenance

In the automobile sector, predictive maintenance involves assessing the status of vehicle subsystems or components, identifying probable problems or defects, and forecasting when repair is necessary. For instance, in a braking system, predictive maintenance may assess sensor data to detect brake pad degradation and notify the driver. Predictive maintenance in automobiles need advanced sensing technology and a strong capacity to analyze data from these sensors to discern patterns and trends in the state of different vehicle components [10]. A sophisticated predictive maintenance model utilizes extensive data sets to process real-time sensor data, assess the vehicle's health, and suggest actions or provide alarms. Predictive maintenance applications for cars used machine learning (ML) techniques, including decision trees, random forests, support vector machines (SVM), neural networks, nearest neighbor algorithms, clustering algorithms, Gaussian processes, naïve Bayes, bootstrapping, and AdaBoost [11]. The majority of vehicle predictive maintenance (VPM) case studies concentrated on the following subsystems: powertrain components, including the engine and transmission; electrical systems and components, including batteries and circuits; and tire pressure and temperature monitoring. The remaining 41% of case studies are allocated among various auxiliary but pertinent sub-systems. The majority of vehicle sub-systems include high-speed rotating machinery and associated components, with most case studies using machine learning for predictive maintenance; hence, the scope has been delineated as evaluating vehicle motion and the wear of machine components [12]. The case studies are comprised of around 50% powertrain, 20% electrical system, and 30% auxiliary system. Figure 2 depict the machine learning in predictive maintenance as shown below.

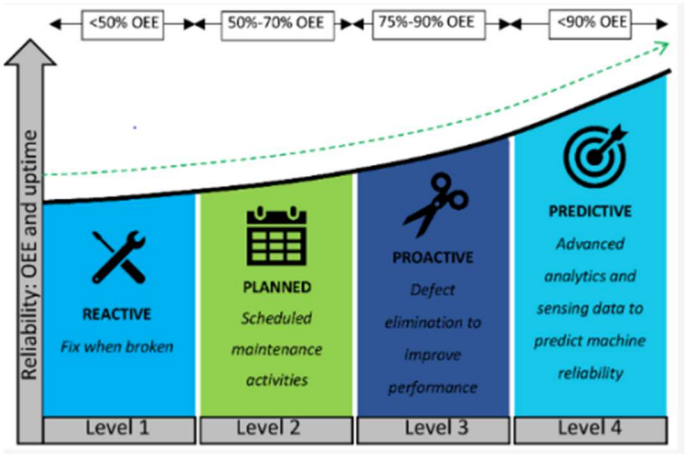

Figure 2. Machine learning in Predictive Maintenance

Despite the prevalent use of telemetry systems, a considerable barrier persists in precisely forecasting vehicle malfunctions. Current methodologies often inadequately address the intricacies and fluctuations of actual driving

situations, resulting in ineffective maintenance regimens. There is a need for more advanced prediction models that can use the extensive data produced by contemporary cars. Conventional maintenance methods often depend on planned inspections or reactive fixes, resulting in inefficiency and high costs. Conversely, predictive maintenance use data-driven methodologies to anticipate the failure of certain components, hence enhancing maintenance scheduling and decreasing total expenses. This research focuses on the development of machine learning models for predicting car component failures with telemetry data.

2. Review of literature

Chaudhuri et al., (2024)[13] studied that connected vehicle fleets have become a crucial element of industrial Internet of Things applications within the context of Industry 4.0 globally. The quantity of cars in these fleets has increased consistently. The use of machine learning algorithms for vehicle monitoring has markedly enhanced maintenance operations. The possibility for predictive maintenance has grown due to the management of equipment via networked smart devices. Benefits are derived from the optimization of uptimes. This has led to a decrease in related time and personnel expenses. It has also yielded a substantial improvement in cost-benefit ratios. This study addressed predictive maintenance issues by using a hybrid deep learning-based ensemble technique (HDLEM) to analyze vehicle fault patterns. The ensemble architecture serving as a predictive analytics engine consists of three deep learning algorithms: modified Cox proportional hazard deep learning (MCoxPHDL), modified deep learning embedded semi-supervised learning (MDLeSSL), and merged LSTM (MLSTM) networks. Sensor data and previous maintenance records are gathered and processed using benchmarking techniques for HDLEM training and evaluation. Modeling and prediction of times between failures (TBF) using multi-source data have been successfully accomplished. The findings acquired are juxtaposed with specified deep learning models. This comprehensive approach has significant promise for enhancing profitability, efficiency, and sustainability in vehicle fleet management systems. This facilitates improved telematics data application, ensuring proactive management towards the intended solution. The superiority of the ensemble approach is shown via various experimental outcomes.

Von Glehn et al., (2024)[14] utilized machine learning methods to ascertain the loading status of trucks—loaded or empty—exclusively using data derived from the vehicle's communication network, particularly from the engine module. They attained an accuracy over 85% for small hauls (0.5 to 5 km) and approximately 95% for long hauls (5 to 500 km). This method enhanced fleet management by minimizing communication between managers and drivers, while also significantly contributing to research on fuel consumption reduction and advanced fault diagnostics. Reducing reliance on inaccurate and delayed human reactions creates opportunities for enhanced research and creative solutions within the transportation sector.

Barapatre et al., (2024)[15] analyzed the essential function of Predictive Maintenance (PdM) in protecting organizations against system malfunctions and incidents. By promptly recognizing and categorizing possible problems, PdM may mitigate accident risks, increase safety protocols, decrease downtime, and optimize vehicle maintenance. Consequently, they used two Machine Learning methods, XG Boost and Logistic Regression, to forecast failure occurrences and identify the optimal framework for certain failure prediction categories. The suggested method is based on the MetroPT dataset and offers several benefits to metro and rail operators. By anticipating maintenance needs, managers may mitigate possible system failures before they escalate into significant problems. This method substantially enhanced vehicle dependability and minimizes downtime in the metro network, hence assuring a more robust and trustworthy transportation system. The data from analog sensors, encompassing air tank pressure, compressor oil temperature, and flowmeter readings, is essential to this framework.

Wang et al., (2023)[16] examined that Predictive Maintenance (PdM) seeks to determine the ideal timing for doing maintenance on an industrial asset based on its current health condition. The objective is to reduce expenses by identifying the ideal moment at which the aggregate of preventative and repair costs is minimized. A data-driven model may forecast the proximity of an asset to a genuine failure, hence facilitating the development of more cost-effective maintenance methods. This study examined survival analysis-based predictive maintenance in the context of Battery Electric Truck (BET) operations. Cox Proportional Hazards and Random Survival Analysis Forest approaches are used for modeling time-to-failure and the corresponding survival functions. Comprehensive telematics data from BET cars in actual operations are used for modeling and analysis. The model's performance is enhanced by feature selection and hyperparameter optimization.

Barber et al., (2023)[17] stated that Electric vehicles (EVs) have the potential to significantly decrease greenhouse gas emissions; but, they provide a challenge for energy distribution infrastructure, which was not originally built to accommodate the increased demand resulting from their widespread adoption. Comprehending the timing of client EV charging and their energy consumption enhances electric utilities' capacity to provide more dependable and cost-

effective electricity to all consumers, facilitating the shift towards sustainable mobility. The research aimed to analyze passenger EV charging data from National Grid's Massachusetts EV Off-Peak Charging Program to ascertain the feasibility of developing generalizable and scalable machine learning models for predicting EV charging energy demand, as well as to identify the minimal geographic granularity for such models. This study introduced a unique approach for estimating charge rates, normalizing charging energy on a per-vehicle basis, considering the inflow and outflow of charging energy within the analyzed system, and using ambient air temperature as a feature variable. Supervised machine learning approaches using random forests were identified as superior for accuracy, complexity, and computing intensity. This study effectively developed and implemented a precise service territory model and highlighted the difficulties of using telematics data for demand modeling.

Moosavi et al., (2023)[18] examined the use of telemetry data and contextual information (e.g., road type, daylight) to characterize a driver's style using tensor representations. Drivers exhibiting similar behaviors are categorized by the grouping of their representations, so creating risk cohorts. Previous at-fault driving incidents and violations function as partial risk indicators. The comparative extent of average recordings (per driver) for each cohort signifies their risk classification, such as low or high risk, which may be assigned to drivers within a cohort. A classifier is then developed using enriched risk labels and driving style representations to forecast driving risk for novice drivers. Empirical research from prominent US cities corroborates the efficacy of this strategy. This method is effective for extensive situations as data may be acquired on a wide scale. Its emphasis on driver-centric risk forecasting makes it relevant to sectors such as automobile insurance. In addition to customized premiums, the framework enables drivers to evaluate their driving conduct in different situations, promoting skill improvement over time.

Chaudhuri et al., (2022)[19] introduced the connected vehicle fleets have often constituted a substantial element of industrial Internet of Things (IIoT) scenarios globally, in accordance with Industry 4.0 guidelines. The quantity of cars in these fleets has increased consistently. The monitoring of these vehicles using machine learning techniques has markedly enhanced the maintenance operations of these systems. In recent decades, the possibility for predictive maintenance has expanded due to the management of equipment via networked smart devices. Predictive maintenance has shown advantages in optimizing uptime. This has led to a decrease in time and personnel expenses related to inspections and preventative maintenance. It has also yielded substantial cost-benefit ratios for commercial earnings. This issue is examined using LSTM Autoencoders to analyze car fault patterns in relation to significant vehicle features. This functions as a predictive analytics engine for this issue. Real-world data gathered from automotive garages is used for the training and testing of LSTM Autoencoders. This approach is contrasted with several support vector machine variations. This approach enables the enhanced use of telemetry data to provide proactive management towards the targeted solution. This method's superiority is shown by many experimental outcomes.

Vanjire et al., (2022)[20] studied the automobile industry employs many electronic components to assess vehicle condition. Data produced by the vehicle component may be used for several purposes, including diagnostics, maintenance, and prognostics (predictive diagnosis). Satisfactory measures have been implemented in the automobile sector for both on-board and off-board diagnostics to facilitate vehicle diagnostics and maintenance efficiency. Given human constraints in expediting analysis and maintenance forecasts, the automation of electronics and data science may provide several solutions, including diverse predictive diagnoses derived from past data. Numerous researchers are now engaged in various sectors of machine learning, with data science yielding superior outcomes in medical predictions. This is also applicable to the automobile industry and its uses. This study enhanced the regression model methodology to forecast clutch state using various factors obtained from the vehicle's CAN bus system and electronic sensors. Diverse supervised machine learning techniques, including support vector machines, logistic regression, decision trees, and polynomial regression, are used. The outcomes derived from these models are evaluated based on their accuracy in predicting the vehicle clutch state.

Markudova et al., (2021)[21] analyzed an extensive fleets of industrial and construction vehicles require regular maintenance procedures. Coordinating these activities may be difficult since the ideal schedule is contingent upon the vehicle's attributes and use. This paper examined a practical industrial example in which a telematics service provider assists fleet managers in coordinating repair activities for about 2000 diverse construction vehicles. The diversity of the fleet and the accessibility of historical data promote the implementation of data-driven solutions using machine learning methods. The study discussed the development of per-vehicle predictors designed to anticipate the subsequent day's utilization level and the time left before the next maintenance. They examined the efficacy of both linear and nonlinear models, demonstrating that machine learning algorithms may effectively capture the fundamental trends characterizing non-stationary vehicle use patterns. They also expressly address the absence of statistics for cars freshly included into the fleet. The findings indicate that access to even a small segment

of previous utilization levels facilitates the identification of cars exhibiting similar use patterns and the opportunistic repurposing of their prior data.

Chen et al., (2021)[22] evaluated that Predictive maintenance (PdM) may enhance industrial efficiency by reducing maintenance costs and increasing production. Predicting remaining usable life (RUL) is a critical problem in predictive maintenance (PdM). The remaining useful life (RUL) of a car may be influenced by several external conditions, including weather, traffic, and topography, which can be analyzed using a geographical information system (GIS). Recently, the majority of academics have undertaken studies on Remaining Useful Life (RUL) modeling using sensor data. Due to the high cost of collecting sensor data, maintenance data is comparatively easier to get. This paper pursued to develop a Remaining Useful Life (RUL) prediction model for automobiles using Geographic Information System (GIS) data via a data-driven methodology. This technique initially included researching a data integration strategy because of the differing data types and sample rates of the maintenance and GIS data. Secondly, the Cox proportional hazards model (Cox PHM) was used to develop the health index (HI) for the integrated data. A deep learning architecture known as the M-LSTM (Merged Long-Short Term Memory) network was developed for HI modeling with integrated data that encompasses both sequential and conventional numeric data. The RUL was ultimately mapped using the expected HI and the Cox Proportional Hazards Model. A comprehensive experimental analysis using an extensive real-world fleet maintenance dataset from a UK fleet firm demonstrated the efficacy of the suggested approach and the influence of GIS parameters on the vehicles examined.

Rahat et al., (2020)[23] introduced the developments in telematics and networking systems have created new prospects for predictive maintenance. The quantity of sensors deployed on vehicles is progressively rising, and manufacturers are seeking innovative methods to enhance fleet uptime while simultaneously minimizing expenses associated with unforeseen failures. The aggregated data from automobiles is sequential, making it pertinent to explore current methodologies for modeling partly observable state sequences to identify prevalent failure patterns. This study presented a novel method for forecasting turbocharger problems in Volvo vehicles. The first phase of the strategy involves simulating a series of readouts from each vehicle using a Markov process. They find the most informative signals and then use spatial similarity clustering to the readouts. They regard each cluster as a Markov state and then transform the history of a truck into a trajectory of states. This trajectory is then linked with repair information to provide a typical sequence labeling issue. Ultimately, they built a hidden Markov model (HMM) classifier to evaluate the equipment's health status. Empirical assessments conducted on the real-world dataset of vehicles indicate that the suggested strategy enhances the AUC score of the final system by up to 6% in forecasting turbocharger failures.

Girma et al., (2019)[24] despite improvements in car security systems, auto-theft rates have escalated over the last decade, and cyber-security assaults on internet-connected and autonomous vehicles are emerging as a novel danger. This study introduced a deep learning model capable of identifying drivers by their driving patterns using car telematics data. The suggested Long-Short-Term-Memory (LSTM) model forecasts the driver's identity by analyzing the distinct driving patterns derived from car telematics data. Since telematics constitutes time-series data, the issue is framed as a time series prediction challenge to leverage the inherent sequential information. The efficacy of the proposed method is assessed using three naturalistic driving datasets, yielding very accurate predictive outcomes. The model's resilience to noisy and aberrant data, often resulting from sensor malfunctions or environmental influences, is also examined. The results indicate that the predictive accuracy of the proposed model is commendable and surpasses other methods, even in the presence of abnormalities and noise within the data.

3. Problem statement

Despite the extensive use of telemetry systems, a considerable barrier persists in precisely forecasting vehicle malfunctions. Current methodologies often inadequately address the intricacies and fluctuations of actual driving situations, resulting in ineffective maintenance regimens. There is a need for more advanced prediction models that can use the extensive data produced by contemporary vehicles.

4. Research Methodology

Figure 3 depicts a typical iterative machine learning workflow, consisting of several key stages. It begins with Data validation to ensure the dataset is accurate and complete. Then, Feature selection identifies the most relevant variables. Route definition outlines the steps for data processing, followed by Dimensionality reduction to simplify the dataset by reducing the number of variables. Next, Data normalization scales the data for consistency, and Training and test sets split the data for model training and evaluation. The Machine learning step involves building

and training the model, and finally, Evaluation of results assesses the model's performance. The methodology for determining the loading state of a truck is composed of nine steps, as described below, and illustrated in Figure 3:

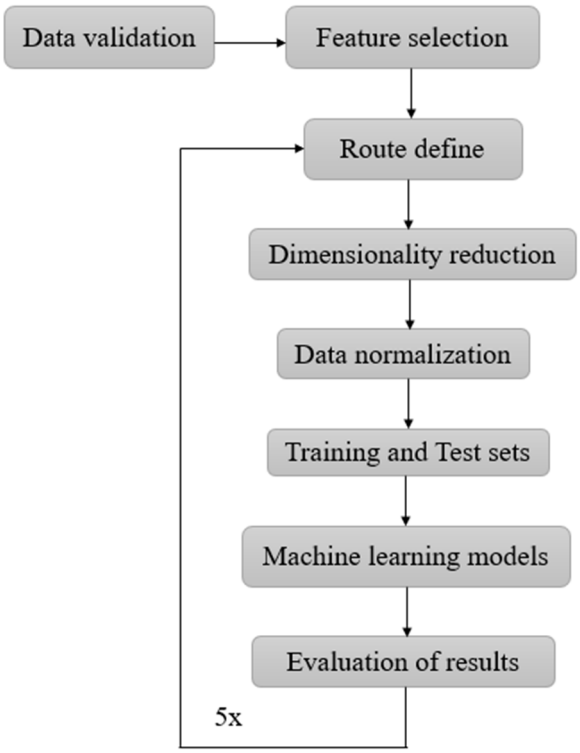

Figure 3. Proposed flowchart

Figure 3 illustrates the validation of the truck's operational data with external sources. The projected gasoline consumption is juxtaposed with the refilling volume recorded by the gas station. Subsequently, the features or variables designated for input into the machine learning algorithm are chosen. The research path is then categorized as short (0.5 to 5 km) or long (5 to 500 km). PCA is used to decrease the dimensionality of the data while maintaining its original variance. PCA begins with two variables and incrementally adds one dimension until the explained variance exceeds 90%. The data is standardized to achieve a mean of 0 and a standard deviation of 1. The dataset is divided into two segments: 75% designated for training and 25% allocated for testing. The machine learning techniques are used on the training set, using all features and the PCA dimensions that account for 90% of the variation. The accuracy of each outcome is assessed on the test set, and the validation process from steps '3' to '9' is conducted ten times for each study route.

4.1 Data validation for volume and path measurement

Three methods were considered for calculating the volume of fuel consumed:

The fluctuation of the fuel tank level (%TL). This approach calculates consumption by multiplying the nominal capacity (NC) in liters of the tank, as indicated by the manufacturer, with the percentage change in the gasoline level (SPN 96) throughout the refueling procedure, as outlined in Equation (1).

$$Vol_{k=TL} = (\%TL_{MAX} - \%TL_{MIN}) * NC \quad (1)$$

The total diesel consumption measured in liters (TFU). This technique employs the data captured by SPN 250 for DAF vehicles and SPN 5054 for Volvo trucks. The estimated consumption is determined by the difference in the values provided by the respective SPN at the end and commencement of the operational state, as outlined in Equation (2).

$$Vol_{k=TFU} = (TFU_{MAX} - TFU_{MIN}) \quad (2)$$

The instantaneous fuel flow rate (FR). This technique uses the value documented by SPN 183. The used value is derived by multiplying the instantaneous fuel flow rate (Q) (SPN 183), expressed in liters per second, by the time interval (t) in seconds, as delineated in Equation (3).

$$Vol_{k=FR} = \sum_{i=1}^{n} \quad Q_{ij}(t_i - t_{i-1}) \tag{3}$$

The first method provides a rough estimate of fuel consumption between fills, the second between runs, and the third at each captured and transmitted data.

Equation (4) delineates the correlation between the distance traveled (D) as shown by the fluctuation of the odometer (ΔO) (SPN 917) and the estimated displacement, which is derived from the product of the recorded speed (V) (SPN 84) translated to km/s and the elapsed time (t).

$$\Delta O_{k=odometer} = O_{MAX} - O_{MIN} \approx D_{K=speedometer} = \sum_{i=1}^{n} \quad V_i(t_i - t_{i-1}) \tag{4}$$

4.2 Inserting the relief among relevant features

The good approximation obtained by calculating the volume based on instantaneous flow (Q) (SPN 183) and the route based on speed (V) (SPN 84) made it possible to include relief features in the list of characteristics available for future analysis [25]. Since the validation of the flow variables makes it possible to calculate fuel economy (FE), as described by Equation (5), considering the relief condition:

$$FE_k = \frac{D_k}{vol_k} = \frac{\sum_{i=1}^{n} \quad V_{ik}(t_i - t_{i-1})}{\sum_{i=1}^{n} \quad Q_{ik}(t_i - t_{i-1})} \tag{5}$$

Where the index "k" represents uphill, downhill, flat or any relief condition. Downhill represent the current altitude is lower than the altitude at the previous instant. Uphill signify the current altitude is higher than the altitude at the previous instant. Flat denotes the current altitude which is equal to the altitude at the previous instant.

4.3 Feature selection

The characteristic group was defined by its creation method. Figure 4 depicts the development of the 29 features.

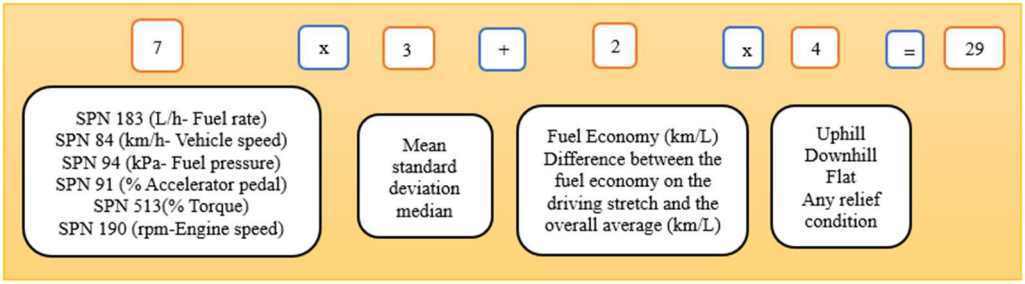

Figure 4. Feature group creation

4.4 Route

Taking on research that conducted trials on 5 km journeys under varying traffic conditions [26], we use this metric to delineate short distances (0.5 to 5 km) from long lengths (5 to 500 km). It is important to note that the supplied dataset contains no distances above 500 km.

4.5 Dimensionality reduction

Principal Component Analysis (PCA) is a robust statistical method used to efficiently decrease the dimensionality of a dataset. It does this by mapping the data to a lower-dimensional space while retaining the maximum amount of inherent variance [27]. PCA does this by finding and extracting the principal components, which represent the primary directions of maximal variability within the dataset. The major components are obtained from the eigenvectors of the covariance matrix of the dataset. The eigenvalues of the covariance matrix measure the degree to which each primary component accounts for the variation in the dataset. The cumulative explained variance, an

essential statistic, is derived from the summation of the squared explained variances, reflecting the total percentage of variation attributed to the chosen principal components.

4.6 Machine learning models

To evaluate the efficacy and enhancement of machine learning algorithms, it is important to partition data into a minimum of two subsets: training and testing. The training data is a portion of the available data used to modify the algorithm's parameters. Test data is a subset of novel, unobserved data used to evaluate the accuracy and generalization of the algorithm on data it has not before seen. Dividing data into training and test sets mitigates overfitting, a condition where the algorithm excels on training data but underperforms on test data, indicating that it has assimilated the unique characteristics of the training data while failing to grasp the overarching patterns. Consequently, partitioning data into training and test sets is an essential procedure in machine learning to guarantee that the algorithm can acquire knowledge from the data and provide dependable predictions on novel data.

The classification process used five machine learning algorithms which is described below.

- **Random Forest:** The Random Forest method is an ensemble learning technique used primarily for classification and regression tasks in machine learning. It constructs multiple decision trees during training and merges them to improve the overall predictive accuracy and control overfitting. Each tree is built from a random subset of data, and at each node, it considers a random subset of features to determine the best split, which introduces diversity among the trees. The final prediction of the model is typically made by averaging the outcomes (in regression) or by taking the majority vote (in classification) from all the individual trees. This approach enhances robustness, reduces variance, and improves the model's generalization ability, making it particularly effective for handling large datasets with complex patterns and interactions.
- **K-NN:** The K-Nearest Neighbors (K-NN) method is a simple, yet effective, supervised machine learning algorithm used for classification and regression tasks. It operates by finding the 'k' closest data points (neighbors) in the feature space to a given input data point and making predictions based on their values. For classification, the input is assigned to the most common class among its 'k' nearest neighbors, while for regression, the output is calculated as the average of the values of its neighbors. K-NN is a non-parametric method, meaning it makes no assumptions about the data distribution, and is highly effective for problems with small to medium-sized datasets.
- **SVM:** Support Vector Machine (SVM) is a supervised machine learning algorithm widely used for classification and regression tasks. The core idea of SVM is to find the optimal hyperplane that separates data points of different classes in a high-dimensional space. It works by identifying the maximum-margin hyperplane that has the greatest distance to the nearest data points (support vectors) from each class, ensuring a robust separation and minimizing classification errors. SVM is particularly effective in high-dimensional spaces and is known for its ability to handle non-linear classification tasks using kernel functions, which transform the input space into a higher-dimensional space where a linear separation is possible.
- **Adaboost:** AdaBoost, short for Adaptive Boosting, is an ensemble learning method used to improve the performance of machine learning algorithms, particularly for classification tasks. It works by combining multiple weak classifiers, which are models that perform slightly better than random guessing, to create a strong classifier with high accuracy. AdaBoost iteratively trains these weak classifiers on various weighted versions of the training data, emphasizing the samples that were previously misclassified. At each iteration, the algorithm adjusts the weights of the samples to focus on the more challenging cases, thus creating a more accurate and robust model. The final model is a weighted combination of all the weak classifiers, where each contributes according to its accuracy, resulting in a powerful predictive model that can handle both linear and non-linear relationships in the data.
- **Logistic regression:** Logistic regression is a statistical method used for binary classification problems, where the outcome or dependent variable has two possible values, such as "yes" or "no," "true" or "false," or "0" and "1." Unlike linear regression, which predicts a continuous output, logistic regression uses the logistic function (or sigmoid function) to model the probability of a given input belonging to a specific class. This function outputs values between 0 and 1, representing the probability of the dependent variable being in a particular class. The method determines the best-fitting model by maximizing the likelihood of correctly classifying the training data.

5. Result and discussion

Table 1 presents a summary of route attributes and their corresponding explained deviations. Significantly, there are a greater number of samples for long routes than for short routes, and explained variances showed similar tendencies in both route groups. This table summarizes the characteristics of two types of routes—short and long stretches—by providing their minimum, average, and maximum distances in kilometers, as well as the explained variance for each. For the short stretch, the minimum route is 0.5015 km, the average is 1.846 km, and the maximum is 5.323 km, with an explained variance of 92.7%. In contrast, the long stretch has a minimum route of 5.00 km, an average of 68.92 km, and a maximum of 253.45 km, with an explained variance of 91.7%. The explained variance indicates how well the data captures the variability in these route lengths.

Table 1. Route characteristics and explained variance

Description	Minimum Route (km)	Average route (km)	Maximum route (km)	Explained variance (%)
Short stretch	0.5015	1.846	5.323	92.7%
Long stretch	5.00	68.92	253.45	91.7%

Table 2 facilitates a comparative analysis of eight classification methodologies, with and without PCA, across both short and long route conditions. From this table 2, the accuracy results of different classifiers—AdaBoost, K-NN (K-Nearest Neighbors), Logistic Regression, Random Forest, and SVM (Support Vector Machine)—applied to both short and long stretches. For the short stretch, K-NN, Random Forest, and SVM achieve the highest accuracy of 95.97%, while Logistic Regression has the lowest at 93.23%. For the long stretch, K-NN again shows the highest accuracy (95.97%), followed by Random Forest at 94.67% and SVM at 94.23%, with Logistic Regression being the lowest at 92.42%. AdaBoost performs slightly better on the short stretch (94.30%) compared to the long stretch (93.6%). To enhance clarity, the most favorable outcomes have been accentuated in black.

Table 2. Accuracy results for different classifier

Description	Adaboost	K-NN	Logistic Regression	Random Forest	SVM
Short stretch	94.30	**95.97**	93.23	**95.97**	**95.97**
Long stretch	93.6	**95.97**	92.42	94.67	94.23

Figure 5 compares the accuracy of five different machine learning models—Adaboost, K-Nearest Neighbors (K-NN), Logistic Regression, Random Forest, and Support Vector Machine (SVM)—for two different data conditions: "Short stretch" (blue bars) and "Long stretch" (orange bars). K-NN and Random Forest models exhibit the highest accuracy for both conditions, with K-NN showing slightly better performance in both short and long stretches. Logistic Regression has the lowest accuracy across both conditions, especially for the long stretch. Overall, the accuracy of most models is higher for short stretch data, except for Adaboost, where the long stretch performs marginally better.

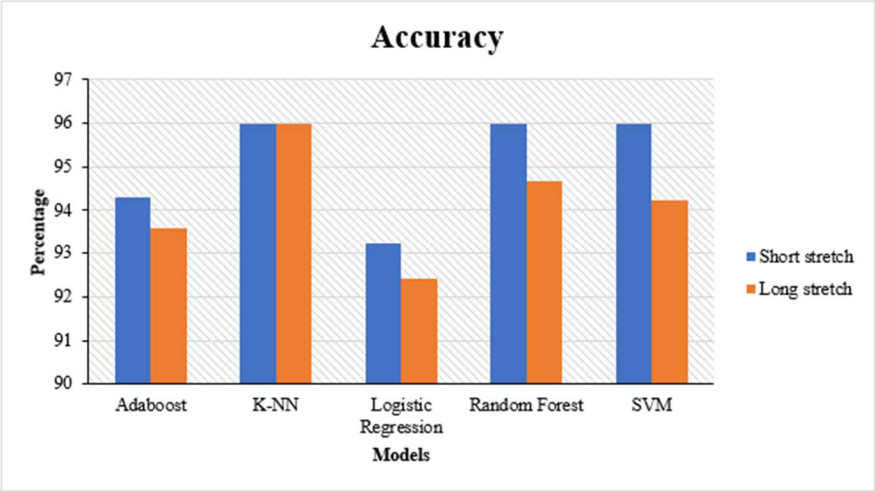

Figure 5. Accuracy of different classifiers

Table 3 depict the fleet information as shown below. The analysis is based on the data generated and insights obtained from four crucial tables: Fleet Information, Sensor Readings, Predictive Maintenance, and Maintenance History.

Table 3. Fleet information

Vehicle ID	Vehicle Type	Mileage
1	**Electric bus**	**45000**
2	**Hybrid car**	**30000**
3	**Electric truck**	**25000**
4	**CNG-powered Van**	**68000**

Figure 6 depicts the Fleet Information on the cars, including their categorization, manufacture year, current mileage, and the date of their most recent repair. This information provides a foundational reference for understanding the initial conditions of the fleet. The analysis indicates that the fleet comprises a diverse range of vehicle types, including electric buses, hybrid cars, electric trucks, CNG-powered vans, and hybrid buses. The diverse range of vehicle types reflects the current trend of environmentally friendly transportation fleets.

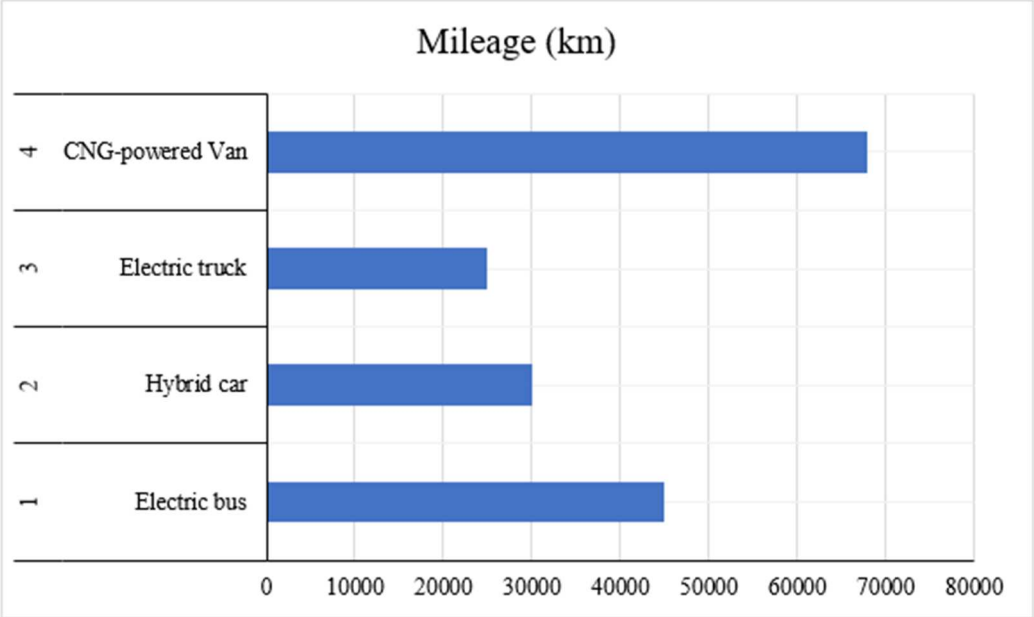

Figure 6. Comparison of Mileage (km)

Table 4 illustrates the Sensor Readings logs real-time data, including engine temperature, battery voltage, and brake wear percentages. This dataset provides a snapshot of the automobiles' state at certain periods. The research uncovers variations in sensor readings, with each vehicle exhibiting unique patterns determined by use and operating conditions. For instance, the electric car has a heightened engine temperature that corresponds to the demands of demanding jobs. The brake wear percentages display diversity, showcasing the system's ability to adapt to different vehicle components.

Table 4. Sensor reading

Vehicle ID	Engine Temperature (degree Celsius)	Battery Voltage (V)	Brake Wear (%)
1	70	350	20
2	82	360	15
3	90	320	35
4	85	330	20

Figure 7 depicts the comparison of sensor readings as shown below. There are four lines in the graph, each representing a different vehicle ID. The x-axis is not labeled but likely represents the distance traveled in some units as it's numbered from 0 to 400. Higher brake wear is correlated with higher vehicle ID as shown below.

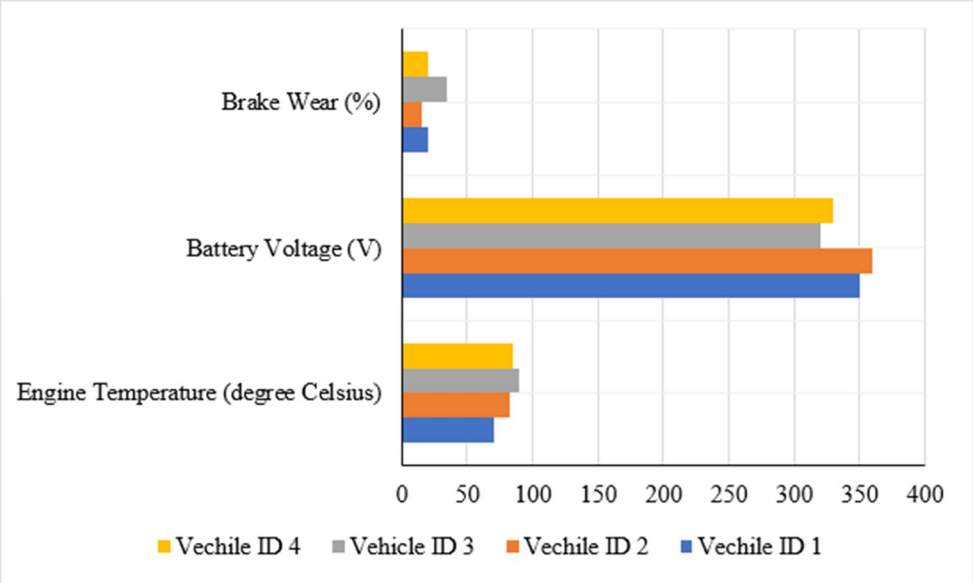

Figure 7. Comparison of Sensor readings

Table 5 displays the Predictive Maintenance data on the suggested intervals for maintenance, the most recent readings of the maintenance odometer, and the future due date for maintenance for each vehicle. The predictive maintenance algorithms modify these plans based on the sensor data received in real time. The investigation shows that the algorithm effectively considers the state of each component, leading to maintenance recommendations that align with the unique characteristics of each vehicle.

Table 5. Predictive Maintenance

Vehicle ID	Maintenance Type	Recommended Interval (KM)	Last Maintenance Odometer Reading (km)
1	Battery Inspection	10,000	40,000
2	Brake system check	5,000	30,000
3	Engine diagnostic	15,000	15,000
4	Transmission Service	20,000	45,000

Table 6 displays the categories of the mileage readings of the maintenance history data for each vehicle. This information provides useful insights into the effectiveness of the predictive maintenance system in identifying and addressing defects. The data reveals a proactive approach, in which maintenance procedures are carried out before catastrophic breakdowns occur. For instance, the electric bus undergoes a battery inspection at 40,000 km, demonstrating the system's ability to proactively address any issues, thereby enhancing the overall reliability of the fleet.

Table 6. Maintenance history

Vehicle ID	Maintenance Type	Odometer Reading (km)	Last Maintenance Odometer Reading (km)
1	Battery Inspection	40,000	40,000

2	Brake system check	5,000	30,000
3	Engine diagnostic	15,000	15,000
4	Transmission Service	20,000	45,000

Engine Temperature Variation: The investigation demonstrates the efficacy of the predictive maintenance system in identifying changes in engine temperature. The percentage change accurately reflects actual fluctuations, showcasing the system's ability to adapt maintenance in response to changing conditions. For instance, if the electric car has a 20% increase in engine temperature, it would promptly recommend a fix to address any issues.

Percentage Change in Battery Voltage: The examination of the percentage change in battery voltage showcases the system's ability to accurately forecast variations. The projected modifications correspond closely with the factual fluctuations seen in the sensor data. The capacity to adapt is essential for maintaining the durability of electric and hybrid cars, since the condition of the battery is a major determinant. The system suggests periodic maintenance tasks to ensure consistent battery performance. The percentage change analysis for brake wear assesses the system's capacity to properly forecast the deterioration of braking systems. The capacity to adjust is crucial in order to provide the best possible brake performance, hence improving both safety and operating efficiency.

The system showcases its capacity to adapt flexibly to different kinds of vehicles and operating conditions, as seen by the fleet data, sensor measurements, and predictive maintenance timetables. The investigation on percentage change validates the system's adaptability, showing its ability to predict variations in crucial parameters. Implementing this proactive method results in reduced times of idleness, cost savings, and improved overall efficiency of the fleet. The maintenance history serves as further verification of the system's effectiveness, shown by the prompt interventions recorded in the maintenance records. The study results enhance the ongoing progress of maintenance strategies in the transportation sector, specifically targeting efficiency, dependability, and ecological responsibility. This, in turn, supports the industry's pursuit of sustainable fleet operations.

6. Conclusion and future scope

The study sought to evaluate data from heavy-duty trucks supplied by a heavy haul trucking firm to identify pertinent features for use as input variables in energy-saving research and other studies in road transport via machine learning. In assessing fuel efficiency and distance traveled, it was intended to corroborate the database using external data from the fueling report and the truck's odometer. Three estimations of fuel consumption volume and two estimates of distance traveled were provided. The minor discrepancies (under 5% for consumption and under 1% for distance traveled) between their results substantiate the data utilized for the estimates: percent fuel level (SPN 96), total fuel consumption (SPN 250 or SPN 5054), instantaneous fuel flow (SPN 183), odometer (SPN 917), and speed (SPN 84). This interpretation was crucial for enhancing the perception of the database's reliability. The validation of the methods for calculating volume via instantaneous fuel flow and route via vehicle speed facilitated the computation of energy consumption in segmented parts. The database's construction may have impeded categorization at short distances; thus, the next idea is to adjust the calculations to ensure that all travel initiates from the long distances while considering short distances. For the short stretch, K-NN, Random Forest, and SVM achieve the highest accuracy of 95.97%, while Logistic Regression has the lowest at 93.23%. For the long stretch, K-NN again shows the highest accuracy (95.97%), followed by Random Forest at 94.67% and SVM at 94.23%, with Logistic Regression being the lowest at 92.42%. The sample base will expand as the minimum distance traveled decreases, perhaps allowing for the determination of the truck's loading status during the first meters of travel. This research shown that freight vehicle telematics data may be effectively examined to formulate studies addressing inquiries related to minimizing energy consumption in freight transportation, hence increasing both operational efficiency and environmental sustainability.

References

[1]. Hao, Zhiqiang, Guojun Li, Lei Chen, and Jingcheng Niu. "Exploration and Practice of Predictive Maintenance Technology in Automobile Factories." In *Society of Automotive Engineers (SAE)-China Congress*, pp. 975-982. Singapore: Springer Nature Singapore, 2023.

[2]. Maillart, Arthur. "Toward an explainable machine learning model for claim frequency: a use case in car insurance pricing with telematics data." *European Actuarial Journal* (2021): 1-39.

[3]. Giobergia, Flavio, Elena Baralis, Maria Camuglia, Tania Cerquitelli, Marco Mellia, Alessandra Neri, Davide Tricarico, and Alessia Tuninetti. "Mining sensor data for predictive maintenance in the automotive industry." In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 351-360. IEEE, 2018.

[4]. Divya, D., Bhasi Marath, and M. B. Santosh Kumar. "Review of fault detection techniques for predictive maintenance." *Journal of Quality in Maintenance Engineering* 29, no. 2 (2023): 420-441.

[5]. Patil, Ravindra B., Meru A. Patil, Vidya Ravi, and Sarif Naik. "Predictive modeling for corrective maintenance of imaging devices from machine logs." In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1676-1679. IEEE, 2017.

[6]. Prytz, Rune, Sławomir Nowaczyk, Thorsteinn Rögnvaldsson, and Stefan Byttner. "Predicting the need for vehicle compressor repairs using maintenance records and logged vehicle data." *Engineering applications of artificial intelligence* 41 (2015): 139-150.

[7]. Aremu, Oluseun Omotola, David Hyland-Wood, and Peter Ross McAree. "A Relative Entropy Weibull-SAX framework for health indices construction and health stage division in degradation modeling of multivariate time series asset data." *Advanced Engineering Informatics* 40 (2019): 121-134.

[8]. Baruah, Pundarikaksha, and Ratna B. Chinnam*. "HMMs for diagnostics and prognostics in machining processes." *International Journal of Production Research* 43, no. 6 (2005): 1275-1293.

[9]. Behrendt, Andreas, Nicolai Müller, Peter Odenwälder, and Christoph Schmitz. "Industry 4.0 demystified–lean's next level." *Retrieved March* 3 (2017): 2017.

[10]. Bian, Wenming, and Mark French. "Coprime factorisation and gap metric for nonlinear systems." In *42nd IEEE International Conference on Decision and Control (IEEE Cat. No. 03CH37475)*, vol. 5, pp. 4694-4699. IEEE, 2003.

[11]. Aravind, Ravi, Chirag Vinalbhai Shah, and Manogna Dolu Surabhi. "Machine Learning Applications in Predictive Maintenance for Vehicles: Case Studies." *International Journal Of Engineering And Computer Science* 11, no. 11 (2022).

[12]. Mandala, Vishwanadham. "From Reactive to Proactive: Employing AI and ML in Automotive Brakes and Parking Systems to Enhance Road Safety." *International Journal of Science and Research (IJSR)* 7, no. 11 (2018): 1992-1996.

[13]. Chaudhuri, Arindam, and Soumya K. Ghosh. "Predictive maintenance of vehicle fleets through hybrid deep learning-based ensemble methods for industrial IoT datasets." *Logic Journal of the IGPL* (2024): jzae017.

[14]. von Glehn, Fabio Ribeiro, Bruno Henrique Pereira Gonçalves, Marlipe Garcia Fagundes Neto, and João Paulo da Silva Fonseca. "Telematics and machine learning system for estimating the load condition of a heavy-duty vehicle." *Procedia Computer Science* 232 (2024): 2616-2625.

[15]. Barapatre, Sarvesh, Omkar Gangurde, Rohit Bibwe, and Shweta Tiwaskar. "Predictive Maintenance Framework for Urban Metro Vehicles." In *2024 5th International Conference for Emerging Technology (INCET)*, pp. 1-5. IEEE, 2024.

[16]. Wang, Hao Luo, Xiaoliang Ma, and Per Olof Arnäs. "A Data-driven Survival Modelling Approach for Predictive Maintenance of Battery Electric Trucks." *IFAC-PapersOnLine* 56, no. 2 (2023): 5999-6004.

[17]. Barber, Adam. "Modeling Passenger Electric Vehicle Charging Demand with Machine Learning Using Telematics Data and Temperature." PhD diss., Massachusetts Institute of Technology, 2023.

[18]. Moosavi, Sobhan, and Rajiv Ramnath. "Context-aware driver risk prediction with telematics data." *Accident Analysis & Prevention* 192 (2023): 107269.

[19]. Chaudhuri, Arindam, Rajesh Patil, and Soumya K. Ghosh. "Predictive maintenance of vehicle fleets using LSTM autoencoders for industrial IoT datasets." In *Big Data Privacy and Security in Smart Cities*, pp. 103-118. Cham: Springer International Publishing, 2022.

[20]. Vanjire, Sachin, and Sanjay Patil. "Analysis of Supervised Machine Learning Techniques for Predicting Vehicle Clutch Status." In *ICCCE 2021: Proceedings of the 4th International Conference on Communications and Cyber Physical Engineering*, pp. 563-577. Singapore: Springer Nature Singapore, 2022.

[21]. Markudova, Dena, Sachit Mishra, Luca Cagliero, Luca Vassio, Marco Mellia, Elena Baralis, Lucia Salvatori, and Riccardo Loti. "Preventive maintenance for heterogeneous industrial vehicles with incomplete usage data." *Computers in Industry* 130 (2021): 103468.

[22]. Chen, Chong, Ying Liu, Xianfang Sun, Carla Di Cairano-Gilfedder, and Scott Titmus. "An integrated deep learning-based approach for automobile maintenance prediction with GIS data." *Reliability Engineering & System Safety* 216 (2021): 107919.

[23]. Rahat, Mahmoud, Sepideh Pashami, Sławomir Nowaczyk, and Zahra Kharazian. "Modeling turbocharger failures using markov process for predictive maintenance." In *30th European Safety and Reliability Conference (ESREL2020) & 15th Probabilistic Safety Assessment and Management Conference (PSAM15), Venice, Italy, 1-5 November, 2020*. European Safety and Reliability Association, 2020.

[24]. Girma, Abenezer, Xuyang Yan, and Abdollah Homaifar. "Driver identification based on vehicle telematics data using LSTM-recurrent neural network." In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 894-902. IEEE, 2019.

[25]. Walnum, Hans Jakob, and Morten Simonsen. "Does driving behavior matter? An analysis of fuel consumption data from heavy-duty trucks." *Transportation research part D: transport and environment* 36 (2015): 107-120.

[26]. Rimpas, Dimitrios, Andreas Papadakis, and Maria Samarakou. "OBD-II sensor diagnostics for monitoring vehicle operation and consumption." *Energy Reports* 6 (2020): 55-63.

[27]. Alpaydin, Ethem. *Introduction to machine learning*. MIT press, 2020.