# Detecting High Fuel Consumption in HDVs with Ensemble of Anomaly Detection Models

Berkay Baris Turan, Emre Genc, Inci Nil Akcig,
Neslihan Goztepe, Mehmet Emin Mumcuoglu, Mustafa Unel
*Faculty of Engineering and Natural Sciences, Sabanci University*, Istanbul, Turkey
Email: munel@sabanciuniv.edu

***Abstract*—This paper presents a machine learning (ML)-based system designed to detect high fuel consumption in heavy-duty vehicles (HDVs) using operational data. The system addresses environmental and efficiency challenges in the transportation industry by precisely monitoring fuel consumption to curb CO2 emissions. An ensemble learning method that integrates unsupervised anomaly detection techniques, including Isolation Forest, Autoencoder, and k-NN Regressor models, is proposed. The anomaly detection results from these models are combined using a weighted majority voting (WMV) approach. This method was tested on a dataset comprising 459 driving records and 14 signals collected from 187 HDVs. Additionally, the Local Outlier Factor (LOF) model was employed to validate the ensemble learning method and identify the root causes of the anomalies. This work contributes to the field of transportation efficiency by offering a novel approach to analyzing fuel consumption in HDVs, thereby paving the way for future advancements in sustainable transportation practices.**

***Index Terms*—Machine Learning, Fuel Consumption, Heavy-Duty Vehicles, Anomaly Detection, Environmental Sustainability**

## I. Introduction

Nowadays, heavy-duty vehicles (HDVs) are commonly utilized in various industries associated with logistics and transportation. Therefore, many manufacturers find it compelling to implement fuel anomaly detection systems for HDVs due to environmental and efficiency issues. High fuel consumption is not only a primary cause for greenhouse emissions but it also leads to drastic manufacturing costs and inconveniences for the customers. The application of machine learning (ML) techniques for fuel consumption prediction and anomaly detection in HDVs has been extensively examined. This review highlights key contributions in this domain, focusing on the ML methods employed and their applications.

Supervised learning techniques, where models are trained on labeled datasets, are commonly used for predicting fuel consumption and detecting anomalies in HDVs. Gong et al. [1] focused on categories of factors that influence FC by implementing and comparing the results of Binary Logistic Regression, BP Neural Network, CART Decision Tree and Random Forest. Bousonville et al. [2] compared the performance of k-NN Regression, Artificial Neural Network (ANN) and Gradient Boost Regression to predict FC in medium and heavy vehicles. Schoen et al. [3] implemented Feed-Forward Neural Network (FNN) on data aggregated over fixed window sizes of distance traveled. Mumcuoglu et al. [4] proposed two models using an ensemble of bagged decision trees where the first one classifies 10-minute driving sections as high or normal fuel consumption, while the second model identifies outlier fuel consumption. Additionally, in [5], Support Vector Machine (SVM) model was implemented along with other traditional supervised models. Adhikary et al. [6] proposed Distributed Nearest Hash (DNH) to address the limitations of k-NN model. Furthermore, Uyanık et al. [7] implemented Ridge and Lasso Regression models for FC prediction of ships. Syahputra [8] applied Adaptive Neuro-Fuzzy Inference System (ANFIS) to predict vehicle FC accurately.

Semi-supervised and unsupervised learning methods are particularly valuable when labelled data is limited. These approaches help detect both known and unknown anomalies. Chen et al. [9] utilized both conventional Autoencoders and Convolutional Autoencoders (CAE) for network anomaly detection. Cheng et al. [10] proposed a novel two-layer progressive ensemble method combining Isolation Forest and Local Outlier Factor (LOF) for efficient outlier detection. Xu et al. [11] presents an enhanced method for data anomaly detection named SA-iForest, a combination of Isolation Forest and Simulated Annealing algorithm.

In addition to supervised, semi-supervised, and unsupervised machine learning models, there exists an advanced method known as Ensemble Learning. Ensemble Learning is a sophisticated machine learning technique that combines multiple base models to create a single best prediction model. Ensemble approaches increase prediction, robustness and accuracy by combining numerous models and generally outperform any single component model [12].

This paper aims to employ the versatile and adaptive capabilities of machine learning (ML) through an ensemble of unsupervised learning techniques, to devise a more encompassing and precise fuel consumption anomaly detection model. The dataset utilized in this study comprises 459 records and 14 signals from 187 vehicles, providing a robust foundation for developing models to detect anomalies. An Ensemble Learning approach, encompassing Isolation Forest, Autoencoder, and k-NN Regressor models, was implemented. These methodologies collectively enabled the detection of anomalies in fuel consumption of the vehicles. The results of the Ensemble Learning model were further interpreted with the implementation of Local Outlier Factor (LOF) model. The architecture of the proposed system is demonstrated in Fig 1.

The organization of the paper is as follows: in Section II, the dataset acquisition, standardization, and signal selection are explained. The anomaly detection methodology is presented in section III. In section IV, the results are discussed. Finally, in Section V, the conclusions are presented.

## II. Data and Preprocessing

### A. Dataset Acquisition

The data were gathered from a cloud-integrated system encompassing various models of heavy-duty trucks. The dataset comprises road trip records for each vehicle, documenting different routes with varying slopes and at different time intervals. Signals were obtained through the sensors embedded in the vehicles. The dataset comprises 459 driving records, each containing 14 signals, collected from 187 heavy-duty vehicles (HDVs). The descriptions and units of these signals are detailed in Table I.

TABLE I: Signal Description

| Signal Name | Signal Description | Unit |
|---|---|---|
| DateTime | Date and time signal is given in ”Serial Date Time” format | (yyyy-MM-dd HH:mm:ss.SSS) |
| VehicleID | ID of the vehicle | (from 1 to 187) |
| HghRslutionTotalVehicleDistance | Total distance traveled by the given vehicle | Meters |
| TachographVehicleSpeed | Vehicle speed | Km per hour |
| EngSpeed | Vehicle engine speed | Rpm |
| ActualEngPercentTorque | Vehicle torque in percentage | Percentage (%) |
| AccelPedalPos1 | Accelerator pedal position | Percentage (%) |
| BrakePedalPos | Brake pedal position | Percentage (%) |
| PCCM_Slope | Road slope | - |
| DStgy_dmRdcAgAct | Adblue consumption indicator | Liter (L) |
| EngOilTemp1 | Engine oil temperature | Degree Celsius (°C) |
| EngCoolantTemp | Engine coolant temperature | Degre Celsius (°C) |
| GrossCombinationVehicleWeight | Gross vehicle weight (including carryload) | Kilograms (kg) |
| EngTotalFuelUsed | Total fuel consumed by the given vehicle | Liter (L) |

### B. Data Standardization

To prepare the dataset for further analysis, linear interpolation was utilized to fill in existing NaN values. Specifically, the interpolation of the Brake Pedal Position signal was handled such that positive values were retained and NaN values were set to zero. Additionally, Positive_Slope and Negative_Slope signals were extracted from the PCCM_Slope signal to examine the independent influences of these two signals on fuel consumption. Each entry in these columns was calculated to retain only positive or negative values, with all others set to zero.

Instead of examining second-by-second deviations in the data, a sliding window approach was used to capture more significant deviations in signals that result in high fuel consumption. A window size of 10 minutes and a stride length of 2 minutes were selected for generating the windows. During the sliding window process, the time series signals were converted into mean and standard deviation values for each signal within the corresponding window. Additionally, the Avg_Fuel signal was calculated to reflect the average fuel consumption per 100 km as follows:

$$\text{Avg\_Fuel} = \frac{\text{Total Fuel Consumption}}{\text{Distance Travelled}} \times 100 \tag{1}$$

Moreover, extreme values were removed using the Interquartile Range (IQR) method. Any points outside the range [Lower Bound, Upper Bound] were eliminated. The Lower Bound and Upper Bound were calculated as follows (with $Q_i$ denoting the $i$th quartile):

$$\text{IQR} = Q_3 - Q_1 \tag{2}$$

$$\text{Lower Bound} = Q_1 - 1.5 \times \text{IQR} \tag{3}$$

$$\text{Upper Bound} = Q_3 + 1.5 \times \text{IQR} \tag{4}$$

Lastly, to ensure better comparability between different signals and to prevent any signal from dominating the models due to its scale, the data used to train the models were scaled. Standard scaling was employed due to the data's approximate normal distribution.

### C. Signal Selection

The selection of signals was based on their high correlation with the Avg_Fuel signal and expert knowledge. The following signals were chosen due to their high correlation with Avg_Fuel: acceleration pedal position mean, brake pedal position mean, positive slope mean, negative slope mean, engine coolant temperature mean, and AdBlue consumption mean. Additionally, tachograph vehicle speed mean, tachograph vehicle speed standard deviation, and gross combination vehicle weight mean were selected based on their reported influence on fuel consumption by experts. In total, ten signals, including the average fuel consumption signal, were selected for use in the models.

## III. Anomaly Detection Methodology

### A. Isolation Forest

Isolation forest is a widely used unsupervised machine learning algorithm for anomaly detection. The algorithm creates an ensemble of isolation trees by recursively partitioning the data. Partitioning is performed by randomly selecting a feature at a time and randomly choosing a split value between minimum and maximum values of the corresponding feature. During the partitioning, each data point travels down the tree as far as it is isolated. The path length is the number of splits required to isolate a data point. Thereafter, path lengths are used for calculating an anomaly score for each data point. The shorter the average path length to isolate a point over all trees, it is more likely to be an anomaly [13]. The anomaly score is then calculated by the following formula:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}} \tag{5}$$

where:

- $s(x, n)$ is the anomaly score for a data point $x$ in a dataset of size $n$.
- $E(h(x))$ is the average path length of point $x$ across all trees.
- $c(n)$ is the average path length of unsuccessful searches in a Binary Search Tree, used for normalization.

Data points with anomaly scores close to 1 indicate anomalies.

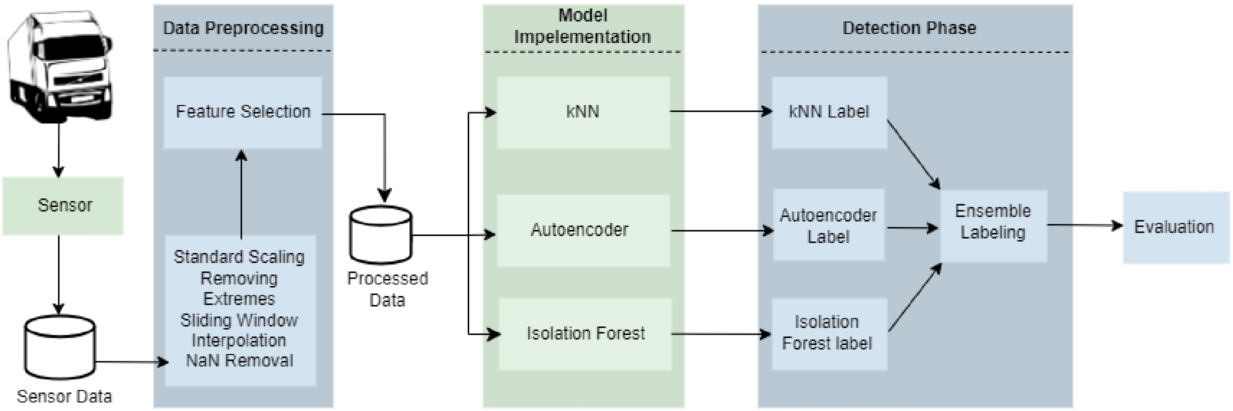

Fig. 1: Architecture of the proposed system

### *B. Autoencoder*

Autoencoders consist of two main parts: the encoder and the decoder. The encoder compresses the input data into a smaller, encoded representation. The decoder then attempts to generate the original data from this encoded representation. The goal is for the autoencoder to learn a representation (encoding) for a set of data, typically for the purpose of dimensionality reduction or anomaly detection [9]. The reconstruction loss, which is minimized by the autoencoder model is calculated as follows:

$$\text{Loss (MAE)} = \frac{1}{N}\sum_{t=1}^{N} \|x_t - \hat{x}_t\| \tag{6}$$

where:

- $N$ is the number of data points.
- $x_t$ is the actual value at time $t$.
- $\hat{x}_t$ is the predicted value at time $t$.

For the purpose of detecting fuel consumption anomalies, an unsupervised basic Autoencoder model is implemented.

### *C. k-NN Regressor*

The k-NN Regressor can be defined as the estimation of the conditional expectation of a variable by calculating the mean of its k nearest neighbors [1]. The Euclidean distance was selected as the distance metric:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{7}$$

where:

- $x$ and $y$ are the feature vectors of the two points.
- $x_i$ and $y_i$ are the $i$-th components of the feature vectors $x$ and $y$.
- $n$ is the number of dimensions (features).

### *D. Ensemble Learning*

Three of the aforementioned models—Isolation Forest, Autoencoder, and k-NN Regressor—employ different methods for identifying anomalies. The Isolation Forest model identifies anomalies as data points that are easy to isolate. The Autoencoder model identifies anomalies as data points that deviate from the common pattern. The k-NN Regressor identifies anomalies as data points with high prediction errors.

To utilize these different methods simultaneously as part of the main prediction model, a simplified version of the weighted majority voting (WMV) algorithm was employed, demonstrated in Fig 2. Since the k-NN Regressor model is the primary prediction model, it was assigned a weight of two, while the Autoencoder and Isolation Forest models were each assigned a weight of one. The WMV label of a data point is calculated as follows:

$$S = \sum_{i=1}^{n} w_i \cdot y_i \tag{8}$$

where:

- $y_i$ is the label (prediction) of the $i$-th model.
- $w_i$ is the assigned weight of the $i$-th model.
- $n$ is the number of predictors.

The final prediction is determined as:

$$\hat{y} = \begin{cases} 1 & \text{if } S \geq 0 \text{ (normal point)} \\ -1 & \text{if } S < 0 \text{ (anomalous point)} \end{cases}$$

### *E. Local Outlier Factor (LOF)*

LOF is used to identify density-based local outliers. The algorithm calculates the LOF score for each point to measure its local deviation from its neighbors. LOF assigns a score to each point based on its density relative to its neighbors. Points with a LOF score of -1 are flagged as outliers, indicating significant deviation in local density compared to their neighbors. These are considered areas of lower density [14]. Points with a score of 1 are considered normal. The signals used in the algorithms are same as the aforementioned selected signals, the number of neighbors is 10 and contamination rate is 0.05.

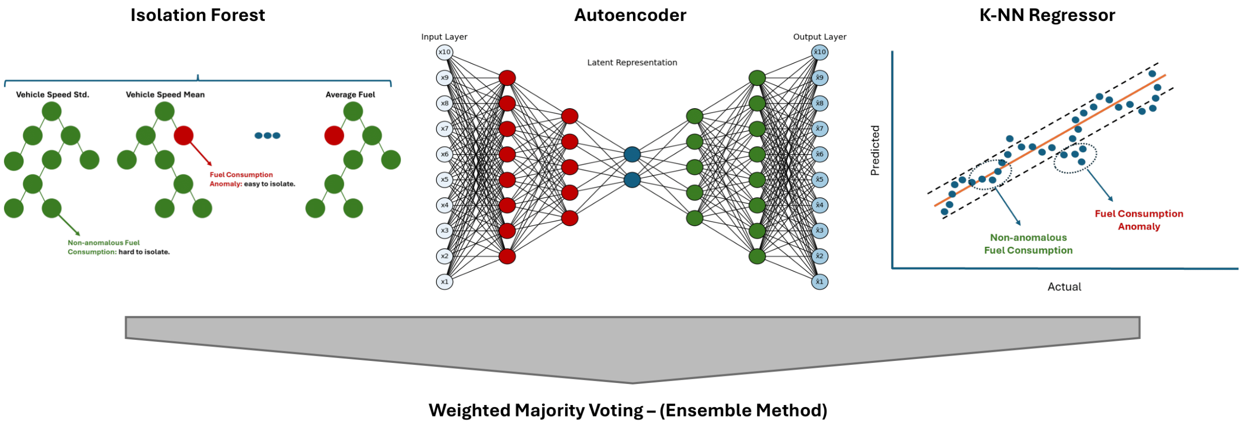

Fig. 2: Structure of the ensemble method.

- Reachability Distance: The reachability distance between two points $p$ and $o$ is defined as:

$$\text{reach-dist}(p, o) = \max(\text{k-dist}(o), d(p, o)) \tag{9}$$

where $d(p, o)$ is the Euclidean distance between $p$ and $o$, and k-dist$(o)$ is the distance of $o$ to its $k$-th nearest neighbor.

- Local Reachability Density (LRD): The local reachability density of a point $p$ is defined as:

$$\text{LRD}(p) = \frac{1}{\frac{\sum_{o \in N_k(p)} \text{reach-dist}(p,o)}{|N_k(p)|}} \tag{10}$$

where $N_k(p)$ is the set of the $k$-nearest neighbors of $p$.

- Local Outlier Factor: The local outlier factor of a point $p$ is then defined as:

$$\text{LOF}(p) = \frac{\sum_{o \in N_k(p)} \frac{\text{LRD}(o)}{\text{LRD}(p)}}{|N_k(p)|} \tag{11}$$

In the model, LOF consolidates findings from three different models and verifies ensemble results to ensure reliable outlier detection.

## IV. Results and Discussion

### A. Anomaly Detection Results

- **Isolation Forest**

In the model used in this paper, input is the combination of all selected signals mentioned above, and the outputs are anomaly scores and anomaly labels. A 5% contamination rate was used, aligning with industry standards, meaning that the 5% of data points with the highest anomaly scores were labeled as anomalies, while all others were labeled as normal points. As a result, the model identified up to 484 anomalous points and 9189 non-anomalous points.

- **Autoencoder**

The inputs used in this model are the selected signals demonstrated above. The output is a reconstructed version of the input data, based on the compressed representation learned by the encoder. Evaluation metric of the model is the reconstruction loss which is MAE. The model has an input layer of 10, consisting of the aforementioned selected signals. The encoder consists of three Dense layers of 8, 5 and 2 units respectively, and each of them relies on the "RELU" activation function. Similarly, the decoder also consists of three Dense layers of 5, 8 and 10 units respectively. However, the third layer utilizes the "Sigmoid" activation function. In total, the model makes use of 298 trainable parameters, while using Adam optimizer, with the aim of minimizing MAE, which is the loss function.

- **k-NN Regressor**

The aforementioned selected signals, except for Avg_Fuel, are used as inputs to predict the output, Avg_Fuel, which represents the average fuel consumption. Using cross-validation, the optimal $k$ parameter (i.e., the number of neighbors) was determined to be 5, resulting in the highest negative squared error. Subsequently, MAE of the actual and predicted values for each data point was calculated, and the 5% of data points with the highest MAE scores were labeled as anomalies, while all others were labeled as normal points. Calculation of MAE score is provided in Equation 6.

Feature importance is expressed through the out-of-bag (OOB) permuted predictor importance, derived from the k-NN model. Essentially, the OOB permuted predictor importance assesses each feature's significance by evaluating the impact of shuffling its values on the OOB data. In this context, features that are more important are expected to have a greater influence.

According to Fig. 4, the most important signal is the positive slope mean, followed by the negative slope mean, vehicle weight mean, and vehicle speed mean. This importance ranking aligns with domain knowledge based on vehicle dynamics and prior research conducted.

Evaluation scores of k-NN regressor model is shown in Table II.

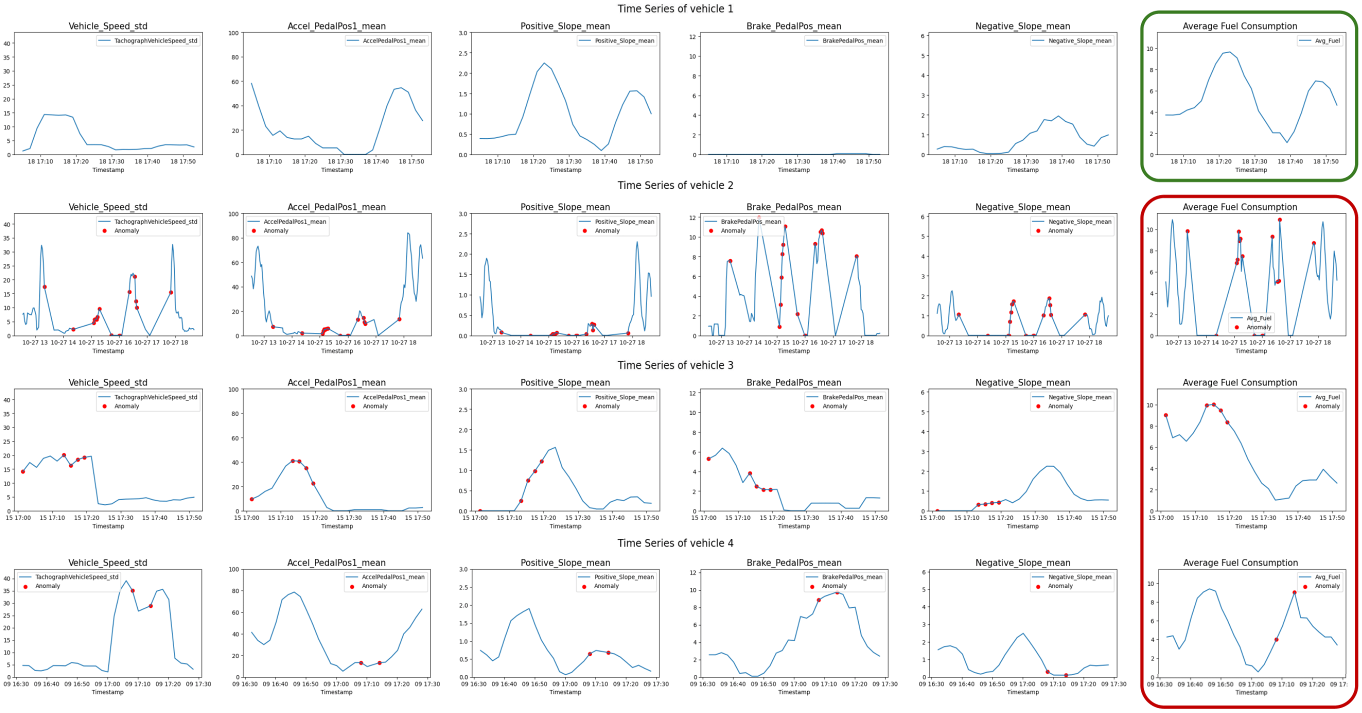

Fig. 3: Time series plots of driving signals and average fuel consumption from sample healthy and anomalous drive cycles of HDVs

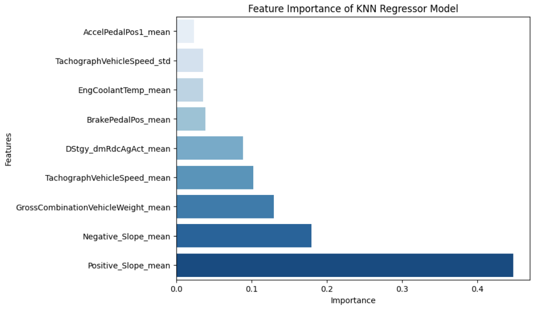

Fig. 4: Feature importance plot for k-NN regressor model

TABLE II: k-NN Regressor Evaluation Scores.

| Evaluation Metrics | | |
|---|---|---|
| MAE | MSE | R-Squared |
| 0.1588 | 0.0596 | 0.9403 |

### B. *Ensemble Method Prediction Results*

In total, 484 data points were labeled as anomalies while 9673 data points were labeled as normal for each model, due to using the same contamination rate (i.e., 5%). After applying the WMV algorithm, 169 data points were identified as anomalous. For further analysis, the ratio of anomalies and anomaly counts were calculated for each vehicle. These vehicles were then sorted in descending order, first by the ratio of anomalies and secondly by the anomaly count.

Fig. 3 shows the time series of signals (i.e., tachograph vehicle speed standard deviation, acceleration pedal position mean, positive slope mean, brake pedal position mean, negative slope mean, and average fuel consumption) recorded during the given cycles of the corresponding signals. Red dots represent the time intervals in which an anomalous high fuel consumption rate is detected by the ensemble model.

The time series of vehicle 1 (first row) represents a typical example of a non-anomalous driving record. As observed, non-anomalous driving cycles exhibit fluctuations in average fuel consumption that resemble fluctuations in the acceleration pedal position mean and positive slope mean. The second row pertains to the vehicle 2 and represents a common scenario where high values and deviations in brake pedal position on a road with a negative slope result in anomalous high fuel consumption rates.

Moreover, the driving cycle of vehicle 3 (third row) exemplifies a case where high acceleration pedal position on a road with a positive slope causes a high fuel consumption anomaly. Timeseries of vehicle 4 (fourth row) illustrates anomalies where a high standard deviation of vehicle speed leads to high fuel consumption. This analysis lead to a classification of cycles that average fuel consumption anomaly occurs under specific cases:

- Case 1: High negative road slope and frequent break pedal use.
- Case 2: High positive road slope and frequent acceleration pedal use.
- Case 3: High standard deviation of tachograph vehicle speed.

Case 1 and Case 2 indicate that high average values of certain combination of signals may lead to an anomaly in FC. On the other hand, Case 3 demonstrates that the driver behavior plays a crucial role regarding FC as the vehicle speed is mostly dependent on a driver's decisions, as observed in the paper [15].

### C. Combining Results of LOF with Ensemble Method

This method uses 169 data points generated by the WMV algorithm. This method seeks to provide a more comprehensive analysis of the causes of anomalies. As illustrated in Fig. 5, the negative slope exhibits the highest number of outliers per signal, followed by the standard deviation of the tachograph vehicle speed, and the average of the gross combination vehicle weight. The feature importance ranking determined by the k-NN algorithm aligns with the findings obtained through LOF method.

The outcomes of the LOF model provide justification for the visual outcomes of the ensemble method. In the first driving cycle of vehicle 106, the high negative road slope and frequent use of the brake pedal cause anomalies in the ensemble learning approach, and LOF also confirms that the average of the negative road slope has the highest anomalies in the data points. Moreover, the anomaly in the driving cycle of vehicle 134 is caused by the high standard deviation of the tachograph vehicle speed. As shown in Fig. 5, the tachograph vehicle speed has the second highest anomaly in the signals.

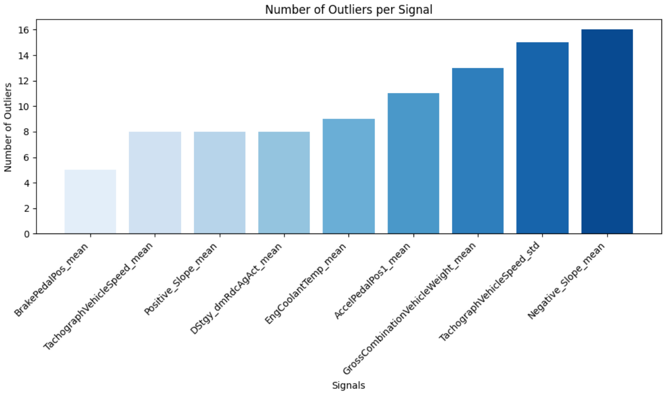

Fig. 5: Frequency plot for identified signal anomalies by Local Outlier Factor (LOF)

## V. Conclusion

This paper proposes a series of machine learning algorithms to address the issue of high fuel consumption in HDVs. The utilization of multiple data standardization methods was vital in terms of increasing the prediction accuracy of the models and obtaining consistent results. The key findings include the identification of anomalous fuel consumption patterns through unsupervised machine learning models. Each of the models followed a different approach for detecting anomalies. The Isolation Forest considers anomalous instances as data points that are easy to isolate. In the Autoencoder model, data points deviating from the common pattern are considered anomalies. The k-NN Regressor on the other hand, pinpoints anomalies as data points having high prediction errors. In order to simultaneously employ these models as a single prediction model, an Ensemble Learning approach with WMV algorithm was adopted. The combination of the LOF method and the WMV algorithm yields results consistent with the outcomes. Both analyses indicate that the average negative road slope and the standard deviation of the tachograph vehicle speed are the primary factors causing anomalies in the dataset.

For future work, expanding the dataset to include more diverse environmental and operational conditions could further validate the model's robustness and generalizability. Explainable AI methods can be utilized to improve the identification of the root causes of fuel consumption anomalies. These methods enable a deeper understanding of the factors contributing to high fuel consumption, thereby facilitating more accurate and actionable insights.

## Acknowledgment

The data and support provided by Ford Otosan is gratefully acknowledged.

## References

[1] J. Gong, J. Shang, L. Li, C. Zhang, J. He, and J. Ma, "A comparative study on fuel consumption prediction methods of heavy-duty diesel trucks considering 21 influencing factors," *Energies*, vol. 14, no. 23, p. 8106, 2021.

[2] T. Bousonville, D. C. Kamga, T. Kruger, and M. Dirichs, "Data driven analysis and forecasting of medium and heavy truck fuel consumption," *Enterprise Information Systems*, vol. 16, no. 6, p. 1856417, 2022.

[3] A. Schoen, R. M. Bagwe, E. C. dos Santos Jr., and Z. Ben-Miled, "A machine learning model for average fuel consumption in heavy vehicles," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 7, pp. 6343–6351, 2019.

[4] M. E. Mumcuoğlu, S. M. Farea, and M. Unel, "Fuel consumption classification for heavy-duty vehicles: a novel approach to identifying driver behavior and system anomalies," in *AEIT International Conference on Electrical and Electronic Technologies for Automotive*, 2023.

[5] A. Bhoraskar, "Prediction of Fuel Consumption of Long Haul Heavy Duty Trucks Using Machine Learning and Comparison of the Performance of Various Learning Techniques," TU Delft, 2019.

[6] S. Adhikary and S. Banerjee, "Introduction to distributed nearest hash: On further optimizing cloud based distributed knn variant," *Procedia Computer Science*, vol. 218, pp. 1571–1580, 2023.

[7] T. Uyanık, Ç. Karatuğ, and Y. Arslanoğlu, "Machine learning approach to ship fuel consumption: A case of container vessel," *Transportation Research Part D: Transport and Environment*, vol. 84, p. 102389, 2020.

[8] R. Syahputra, "Application of Neuro-Fuzzy Method for Prediction of Vehicle Fuel Consumption," *Journal of Theoretical & Applied Information Technology*, vol. 86, no. 1, 2016.

[9] Z. Chen, C. K. Yeo, B. S. Lee, and C. T. Lau, "Autoencoder-based network anomaly detection," in *2018 Wireless Telecommunications Symposium (WTS)*, 2018.

[10] Z. Cheng, C. Zou, and J. Dong, "Outlier detection using isolation forest and local outlier factor," in *Proceedings of the Conference on Research in Adaptive and Convergent Systems*, 2019, pp. 161–168.

[11] D. Xu, Y. Wang, Y. Meng, and Z. Zhang, "An improved data anomaly detection method based on isolation forest," in *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*, vol. 2, 2017, pp. 287–291.

[12] Y.-W. Chung, B. Khaki, T. Li, C. Chu, and R. Gadh, "Ensemble machine learning-based algorithm for electric vehicle user behavior prediction," *Applied Energy*, vol. 254, p. 113732, 2019.

[13] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in *2008 Eighth IEEE International Conference on Data Mining*, IEEE, 2008.

[14] E. Müller, P. I. Sánchez, Y. Mülle, and K. Böhm, "Ranking outlier nodes in subspaces of attributed graphs," in *IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, 2013.

[15] T. Hlasny, M. P. Fanti, A. M. Mangini, G. Rotunno, and B. Turchiano, "Optimal fuel consumption for heavy trucks: A review," in *IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*, 2017.