Journal of Internet Services and Applications, 2026, 17:1, doi: 10.5753/jisa.2026.7108

Failure Profile Characterization in Heavy-Duty Trucks Using Fuzzy Clustering of Telemetry and Repair Data

Juliana Viscenheski [**Univ. Tecnológica Federal do Paraná** | ***juliana.viscenheski@gmail.com***]
Ricardo Lüders [**Univ. Tecnológica Federal do Paraná** | ***luders@utfpr.edu.br***]
Heitor S Lopes [**Univ. Tecnológica Federal do Paraná** | ***hslopes@utfpr.edu.br***]
Thiago H Silva [**Univ. Tecnológica Federal do Paraná** | ***thiagoh@utfpr.edu.br***]

Universidade Tecnológica Federal do Paraná (UTFPR), Av. Sete de Setembro, 3165, Curitiba, PR, 80230-901, Brazil.

Received: 04 December 2025 • **Accepted:** 14 April 2026 • **Published:** 02 July 2026

Abstract Among transportation companies, the most expensive operational costs are due to the maintenance of truck fleets. Data-driven approaches that leverage telemetry and repair data can be effective in addressing such problems. Telemetry data is collected from sensors that monitor truck operational conditions, while repair data is recorded during both scheduled and emergency maintenance. This work characterizes the profiles of Euro 6 heavy-duty trucks using fuzzy clustering to identify vehicles with component failure, specifically within the powertrain system (engine/transmission assembly). Feature selection is based on correlation analysis and extracted from the time series of sensors. Three methods are used to characterize failure profiles: (i) a baseline strategy using averaged feature vectors of failed and non-failed trucks; (ii) Fuzzy C-Means (FCM); and (iii) Fuzzy Self-Organizing Maps (FSOM). These profiles are then used to compute failure risk scores for each truck, based on the similarity to failed and healthy truck references. Our contribution includes a detailed evaluation of FCM and FSOM parameters and their impact on failure results, as well as the identification of usage-related features as stronger predictors of failure than component-specific variables. Results show that the clustering-based strategy can significantly improve failure identification when compared to the baseline. This finding demonstrates the potential to support targeted preventive maintenance with reduced false positives.

Keywords: Predictive maintenance, telemetry, warranty data, clustering, heavy-duty trucks, failure prediction

1 Introduction

Efficient cargo transportation plays an essential role in sustaining industrial growth and competitiveness in Brazil. Although multimodal integration has advanced in recent years, road freight transport continues to dominate the national logistics chain, accounting for nearly 60% of all cargo movement [Tagliatti *et al*., 2024]. Among operational expenditures, variable costs, such as fuel, maintenance, tires, and tolls, remain the primary factors shaping freight pricing structures [da Penha Araujo *et al*., 2013]. Consequently, numerous studies and industrial initiatives have aimed to mitigate vehicle maintenance costs, thereby improving operational efficiency and profitability.

From the perspective of Original Equipment Manufacturers (OEMs), maintenance-related costs are not only a financial issue, but also a matter of brand image and customer satisfaction. Unexpected component failures during the warranty period can increase service costs, reduce fleet availability, and negatively impact customer trust [Khoshkangini *et al*., 2020]. In some cases, higher-than-expected failure rates can result in extended vehicle downtime due to component shortages.

The present work presents a data-driven predictive approach for identifying heavy-duty vehicles with an elevated likelihood of component failure. It advances prior work [Viscenheski *et al*., 2025] by incorporating FSOM clustering, systematically evaluating membership thresholds, and expanding the analysis across two distinct applications. Telemetry data from Euro 6 trucks is combined with historical warranty records to analyze failure patterns in a powertrain subsystem (encompassing engine and transmission elements), which typically accounts for nearly half of Predictive Maintenance (PdM) activities in the heavy-vehicle sector [Ravi *et al*., 2022]. Due to confidentiality agreements, the specific component under analysis cannot be disclosed.

The proposed methodology begins with data preprocessing to ensure operational comparability across vehicles. A correlation-based feature selection is performed, followed by time-series feature extraction to derive relevant temporal and statistical indicators. The analysis compares three classification strategies: (i) a baseline model based on aggregated statistics from failed and non-failed vehicles, and (ii) a clustering-based approach using Fuzzy C-Means (FCM) to identify similarity patterns. (iii) a clustering-based approach using Fuzzy Self Organizing Maps (FSOM) to identify similarity patterns. By constructing representative feature profiles ("signatures") for both groups, the framework enables the identification of vehicles exhibiting behavioral characteristics similar to those observed in previous failures.

The main contributions are twofold: i) an evaluation of how FCM/FSOM clustering parameters affect predictive performance; and ii) an assessment of the most informative features for detecting early signs of component degradation.

The remainder of the study is organized as follows. Section 2 reviews related work. Section 3 describes the data set and details of the methodology. Section 4 presents the results,

followed by a discussion in Section 5. Finally, Section 6 concludes the work.

2 Related Work

According to Mattos *et al*. [2023], PdM enables timely interventions, such as repairs or replacements, minimizing costs while maximizing uptime and operational efficiency. It involves monitoring the condition of vehicle subsystems or components, diagnosing potential faults, and predicting the optimal time for maintenance. Earlier PdM efforts were primarily based on statistical techniques, such as those applied to warranty claim prediction [Wasserman, 1992; Singpurwalla and Wilson, 1998]. However, with the advent of IoT and Industry 4.0, AI-driven approaches have gained importance [Samatas *et al*., 2021]. Data-driven approaches, which are the focus of this work, utilize sensor data and machine learning to detect anomalies and predict maintenance needs without requiring additional knowledge of mechanics. This strategy is especially effective when both abundant historical and real-time data are available [Schlechtingen and Santos, 2011].

Recent studies have increasingly employed machine learning to analyze large-scale datasets [Surucu *et al*., 2023]. For instance, Cao *et al*. [2022] and Mattos *et al*. [2023] apply traditional supervised learning methods to predict failures using vehicle-specific features. Similarly, Principi *et al*. [2019] proposes an unsupervised deep autoencoder for electric motor fault diagnosis, while Shaowu *et al*. [2020] showed promising results in fault classification of hydraulic pumps using a convolutional neural network. A complementary study by Khoshkangini *et al*. [2020] compares two machine learning approaches: autoregressive failure modeling and aggregated individual predictions for heavy-duty vehicles. Such work found that aggregated models outperform regression techniques when data is abundant, whereas regression remains suitable for newer vehicles with limited data.

Heavy-duty trucks pose distinct challenges due to their highly variable operating conditions. One of the most critical aspects of analyzing their performance is understanding the ”application.” It is a term encompassing factors such as cargo weight, road characteristics, and traffic conditions. In this context, Seixas *et al*. [2023] has demonstrated that such operational factors can be effectively clustered to define a vehicle’s application consistently by incorporating slope, vehicle speed, and Gross Combination Weight (GCW), which is the total weight of the vehicle and the goods carried. The authors introduced the load factor metric, which represents the percentage of the distance traveled under full load. The dataset used in their study includes information on cargo classification, road type, traffic profile, and load factor, offering a comprehensive picture of vehicle operations.

Among the works related to the present paper on unsupervised learning, Amruthnath and Gupta [2018] evaluates clustering algorithms, including k-means and hierarchical clustering, for fault detection based on vibration data from exhaust fans, assessing both accuracy and robustness. In Azzaoui *et al*. [2019], FCM clustering was applied to identify anomalies in a continuous distillation system. However, they are concerned with exhaust fans and distillation columns, rather than PdM in vehicles. In a previous work [Viscenheski *et al*., 2025], we proposed a data-driven approach integrating telemetry and warranty data from Euro 6 heavy-duty trucks to identify vehicles with a higher likelihood of powertrain component failure. The study also shows that usage-related features are stronger predictors of failure than component-specific variables, and that clustering-based strategies significantly improve failure identification compared to baseline methods.

3 Data and Methods

The steps followed in the proposed methodology are shown in Figure 1. The operating scenarios or applications are described in Section 3.1, the datasets in Section 3.2, and the clustering approaches in Section 3.3.

Figure 1 illustrates the overall methodological workflow adopted in this study. The process begins with data preprocessing steps. First, the dataset is filtered by vehicle’s application (Step 1). Next, for each vehicle, the variables are computed considering the X days preceding a component replacement or data collection event (Step 2). In Step 3, highly correlated features (correlation > 0.9) are removed to reduce redundancy and multicollinearity. Step 4 excludes vehicles with fewer than N days of available data to ensure statistical consistency. Finally, Step 5 performs time-series feature generation and data cleaning (using TSFRESH), preparing the dataset for modeling.

After preprocessing, 2 alternative analytical strategies are evaluated: a Baseline approach and a Clustering-based approach. In the baseline pipeline, failed and healthy vehicles are separated using K-Fold Cross-Validation. Signatures are defined as the mean of the feature arrays for each group, and distance metrics are calculated between each vehicle and the corresponding signatures. Vehicles are then ranked, and the final performance metric (FB) is computed. In the clustering-based pipeline, FCM/FSOM clustering is used to identify the latent structure of the data. A minimum membership degree is defined to assign vehicles to clusters. Clusters are then labeled as failed or healthy, and cluster centroids are used to define the signatures. As in the baseline approach, distance metrics are computed, followed by ranking and FB evaluation. Together, these steps provide a comparison between a traditional supervised baseline and a clustering methodology for failure detection and ranking.

3.1 Selected Applications

An application is characterized by a combination of engine power, axle ratio, GCW classification, speed classification, and route slope classification. It defines a particular vehicle’s operating configuration, which we will call an application, as shown in Step 1 of Figure 1. Following the criteria presented by Seixas *et al*. [2023], four vehicle applications are selected as shown in Table 1. For confidentiality reasons, applications and their characteristics are represented by letters and numbers.

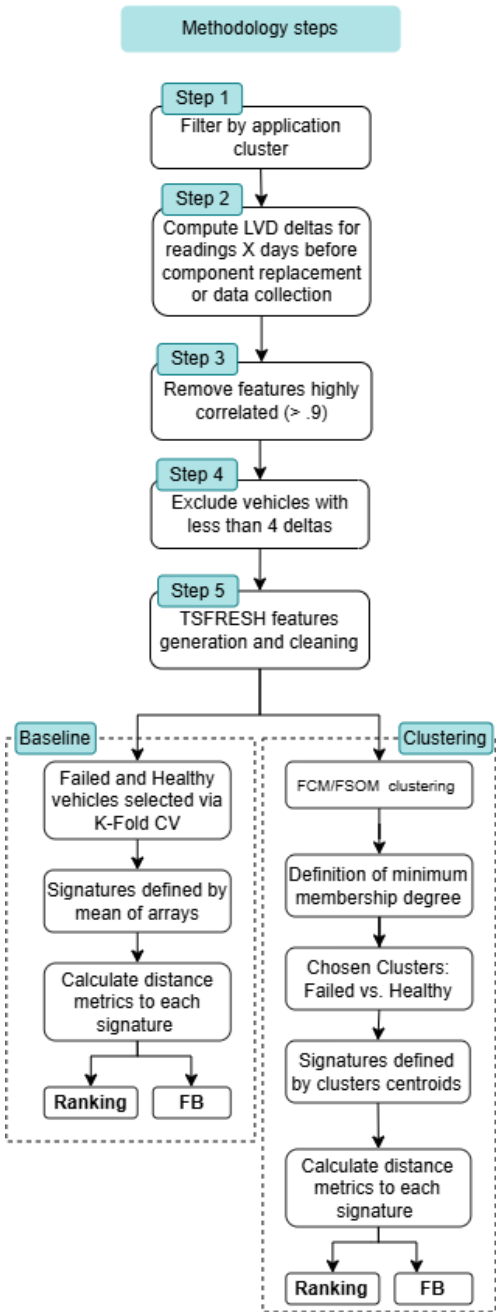


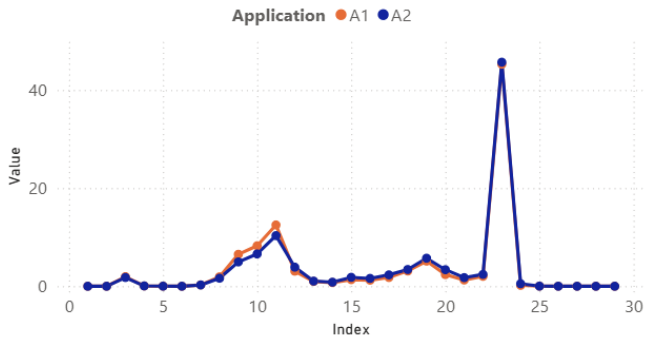
Figure 1. Overview of the methodological workflow.

The GCW information is based on a two-dimensional vector containing 29 weight distribution ranges, from 0–3.5 tons up to over 200 tons. Figure 2 presents the average GCW curve for vehicles of applications A and B with load intervals on the x-axis and the percentage of distance traveled in each interval on the y-axis. Note that average carried weights are substantially different, and their intervals are hidden due to confidentiality.

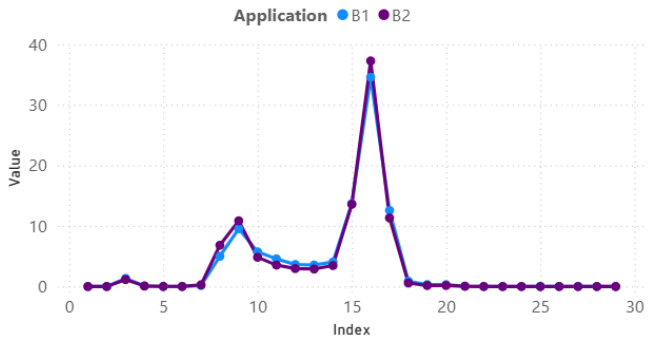
The Speed column is obtained from a distribution of 20 speed ranges. The first range corresponds to 0–5 km/h (speed level 1), while the last represents speeds above 119 km/h (speed level 20). Unlike GCW, the speed distribution uses engine operating time as the aggregation basis. Figure 3 shows the average speed curve for applications A and B with speed intervals on the x-axis and the percentage of time spent in each interval on the y-axis. A significant difference can be observed between the two groups, A and B, with B showing

Table 1. Selected Applications

App Id	Engine	Axle ratio	GCWR	Speed	Slope
A1	A	B	0	10	0
A2	A	B	0	8	2
B1	C	D	2	3	2
B2	C	D	2	9	2



(a) GCW for A1 and A2



(b) GCW for B1 and B2

Figure 2. GCW patterns for applications A1, A2, B1, and B2

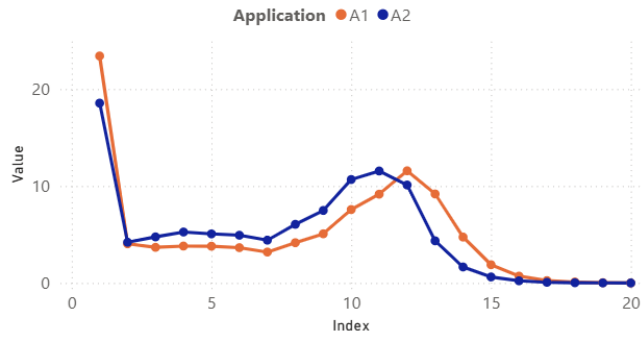
higher speeds than A, and B1 presenting the highest average speed. The speed intervals are hidden due to confidentiality reasons.

The slope classification is built similarly to GCW, using the traveled distance. The distribution comprises 32 ranges, varying from -20% to +20% of road slope at the time of measurement. Figure 4 presents the average slope curve for applications A and B with the road slope intervals on the x-axis and the percentage of distance traveled in each interval on the y-axis. The slope intervals are hidden for confidentiality reasons. It is possible to observe differences between applications A and B, especially between A1 and A2. Although B applications do not show significant variations in slope, they differ in speed and slightly in GCW, which justifies their individual analysis.

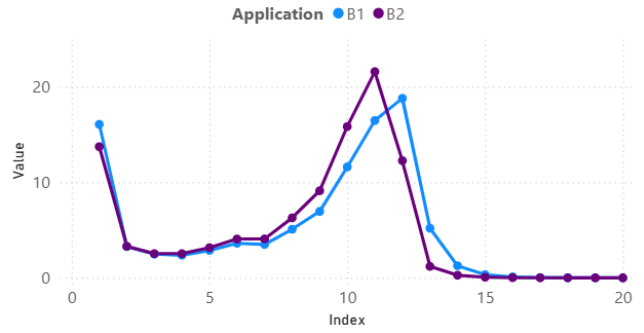
The classification process of GCW, speed, and slope distributions was originally developed by Seixas *et al*. [2023]. Due to the relevant differences between applications A1 and B1, they are considered in this work.

3.2 Datasets and Features

Data used in this study are anonymized, and come from the same fleet of trucks. They result from the combination of two datasets: (i) Logged Vehicle Data (LVD), and (ii) Repair Data. The LVD dataset includes 125 sensor variables

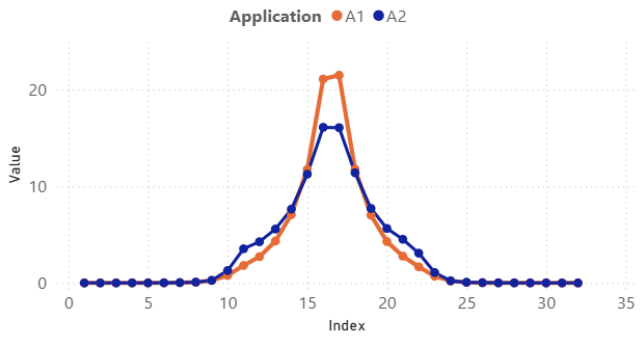


(a) Speed for A1 and A2

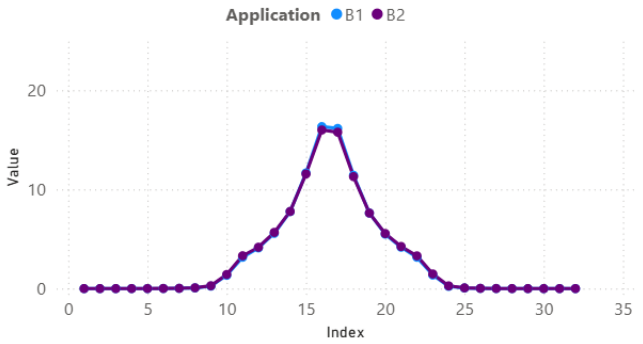


(b) Speed for B1 and B2

Figure 3. Speed patterns for applications A1, A2, B1, and B2



(a) Slope for A1 and A2



(b) Slope for B1 and B2

Figure 4. Slope patterns for applications A1, A2, B1, and B2

collected weekly from monitored vehicles. The initial dataset contains 28,021 vehicles. However, this database aggregates vehicles operating under many different configurations and applications. Since the objective of this study is to analyze failure patterns within specific operational contexts, the dataset was filtered to retain only vehicles belonging to the applications selected in Section 3.1. After applying this filtering step, the resulting dataset contains 1,117 vehicles. This reduction reflects the selection of a homogeneous subset of vehicles operating under comparable application conditions rather than the exclusion of data due to preprocessing constraints.

It is important to emphasize that all variables are normalized based on the total distance traveled or total engine operating hours, allowing for comparison between newer vehicles and those produced several years ago. The repair data (records of component replacements performed at dealerships) include only replacements made within the first two years of vehicle operation, except in cases where additional maintenance plans are in place. From 125 variables, 88 remained after filtering for highly correlated ones. Between two highly correlated variables, for example, the one with higher correlation with failures was kept.

According to Step 5 of Figure 1, feature extraction is performed using the TSFRESH library (a Python tool for time-series analysis). TSFRESH generates features from the time series, such as statistical measures (mean, variance), frequency components (Fourier coefficients), and pattern-based attributes (autocorrelation, entropy), according to Fulcher and Jones [2014]. It is important to note in Step 4 of Figure 1 that a minimum requirement for using TSFRESH is removing vehicles with little information.

The 88 remaining variables are then expanded into a total of 62,806 extracted features. These new features do not directly represent the raw time series but rather statistical and structural characteristics derived from time windows for each vehicle, such as mean, entropy, and dominant frequency, among others. During the extraction process, some features may contain missing values for various reasons. For example, certain combinations of statistical functions and time-series structures may be non-computable when data exhibit low variability, constant values, or short segments within the analysis window. Therefore, features with more than 10% missing values are discarded, reducing the total to 1,173 features. Since these features already represent fixed statistical summaries, the resulting dataset is treated as a static table (one row per vehicle) with columns representing the derived attributes, thereby eliminating the need to consider the original temporal structure.

For the remaining missing values, imputation is performed using the k-Nearest Neighbors (KNN) algorithm with $k = 5$, replacing TSFRESH's default method. Unlike zero-filling or arbitrary numeric substitution, the KNN approach estimates plausible values based on the similarity among vehicles, thereby better preserving the dataset's statistical structure.

For vehicles with a failure event, the LVD readings considered correspond to those recorded up to 120 days prior to the component replacement, as shown in Step 2 of Figure 1. For vehicles without failures, all readings registered within the 120 days preceding the start of the present study are used.

A time window in days is adopted instead of mileage or engine operating time due to the high regularity of use among the vehicle types analyzed, which results in a strong correlation between these variables. The 120-day value is defined based on evidence from the literature and practical knowledge from OEMs and dealerships. Once we have one single failure type, a fixed window is an appropriate approach.

Previous studies, such as Khoshkangini *et al.* [2020], eval-

uated the use of shorter temporal windows. However, the results obtained were below expectations in terms of predictive performance. Extending the window to 120 days aims to provide not only more accurate predictive results but also a longer response time for OEMs in cases of potential quality issues. Additionally, this choice facilitates the alignment of preventive maintenance actions with the pre-scheduled service calendar of dealerships, avoiding the need for additional vehicle trips to workshops. Although a fixed window was adopted for consistency across failure cases, future work may explore adaptive windows or attention-based temporal models that dynamically capture degradation patterns.

The LVD variables represent cumulative values over the vehicle's lifetime. To enable the analysis of behavioral patterns preceding failures, cumulative values were transformed into differences between consecutive readings (for example, consecutive readings of 151,135 and 151,385 km result in $\Delta = 250$ km).

Due to the high correlation among sensors in the powertrain system, Step 3 of Figure 1 considers a Pearson correlation coefficient equal to or greater than 0.7. It is used to remove redundant variables and preserve those most associated with failures. More restrictive thresholds (0.8 and 0.9) were also tested, but a value of 0.7 provides a better balance between dimensionality reduction and preservation of informative variables. The correlation analysis also aims to reduce the feature space, thus helping to mitigate the curse of dimensionality.

It is important to highlight that the dataset presents a natural class imbalance, with the number of failed vehicles being substantially smaller than the number of non-failed vehicles in both applications (see Table 2). This imbalance reflects real-world fleet conditions, where component failures are relatively rare events compared to normal operation. Such an imbalance increases the difficulty of the predictive task, as models may be biased toward the majority class. Therefore, performance metrics based on ranking (P@N) and failure-rate analysis at probability thresholds were adopted to provide a more informative evaluation than overall accuracy, which could be misleading under highly imbalanced conditions.

Table 2. Number of vehicles in each application after applying the TSFRESH filter

Application	Non-failed Vehicles	Failed Vehicles
A1	735	114
B1	220	48

3.3 Baseline and Clustering Approaches

Baseline: This approach does not employ clustering algorithms or supervised models. Instead, it performs a direct comparison between the test vehicles and two reference groups: vehicles with failure (group AF) and vehicles without failure (group AS). Using k-fold cross-validation, the mean of the attributes for the failed and non-failed vehicles in the training set is computed, forming the AF and AS signatures, respectively. Each vehicle in the test set is then compared with these two signatures using different distance metrics (Euclidean, Manhattan, and Cosine). Each vehicle is assigned to the group whose signature has the smallest distance, allowing the assessment of its similarity to the failure and non-failure profiles. This approach serves as a reference for comparison with other techniques, enabling evaluation of how well the extracted features distinguish between the two groups. The proposed baseline was intentionally designed to be simple and aligned with the problem structure. By representing each class with its centroid (mean signature), the method defines an easy-to-interpret decision boundary based on feature similarity. This allows us to assess if the extracted telemetry variables alone provide sufficient separability between failed and non-failed vehicles, without the influence of more complex assumptions. Besides, centroid-based distance comparison is computationally efficient and widely used in similarity-based diagnostics, making it an appropriate lower-bound benchmark for evaluating the improvements achieved by fuzzy clustering methods.

Clustering: The Fuzzy C-Means (FCM) and Fuzzy Self-Organizing Map (FSOM) algorithms are used to group vehicles based on their similarity, together with a soft membership assignment to handle uncertainty and borderline cases. The number k of clusters is determined based on the metrics of Sum of Squared Errors (SSE), Silhouette, and Calinski–Harabasz. Although these metrics (particularly Silhouette) assume approximately spherical cluster shapes, which may not perfectly reflect the true data structure, the final number of clusters is also determined by visual inspection of the metric curves. The goal is to achieve a balance between good group separation (high Silhouette and Calinski–Harabasz values) and low reconstruction error (decreasing SSE). Values of k that demonstrate consistency across criteria and provide a coherent interpretation of the cluster structure are selected, thereby avoiding both over- and undersegmentation of the data.

It is important to note that clustering is performed before defining the test set, since the objective of this study is to explore latent behavioral patterns and characterize failure signatures rather than to build a purely predictive classifier. While this design may introduce some risk of information leakage in a strict predictive evaluation setting, the baseline method based on k-fold cross-validation provides a complementary evaluation with explicit training–test separation.

After clustering vehicles, the cluster with the highest proportion of failed vehicles is labeled AF, while the one predominantly composed of non-failed vehicles is labeled AS. Vehicles with low membership degrees or belonging to highly heterogeneous clusters are considered part of the test group. Similar to the baseline approach, these vehicles are then compared with the centroids of clusters AF and AS using different distance metrics. The results are evaluated with *Precision@N* and *Failure Probability* metrics as described below.

Precision@N (P@N): Quantifies the proportion of failed vehicles among the N vehicles most similar to the failure signature, according to the selected distance metric. For example, P@10 evaluates the 10 vehicles closest to the failure signature. A value of 1 indicates that all N most similar vehicles experienced a failure.

In addition to the quantitative results, it is important to consider the advantages and limitations of different cutoff levels for the P@N metric. More restrictive cutoffs, such as P@10, tend to prioritize vehicles with the highest similarity to the failure profile, offering higher precision and fewer

false positives. This is especially valuable in contexts where preventive intervention capacity is limited and maintenance costs are high, making it a primary metric for comparing approaches. Conversely, broader cutoffs such as P@50 and P@100 favor greater coverage, potentially capturing more vehicles at risk, albeit at the cost of lower precision and more false positives.

Failure Probability (FP): This metric is defined in this study as a measure of how close a vehicle is to the failure signature and how far it is from the non-failure signature, as shown in Equation (1):

$$\mathrm{FP} = \frac{\mathrm{AS}}{\mathrm{AF} + \mathrm{AS}} \tag{1}$$

with distance AF to the failure signature, and distance AS to the non-failure (healthy) signature.

Higher values of FP (close to 1) indicate greater similarity to the failure pattern, while lower values indicate a behavior more consistent with the healthy pattern. FP is thus interpreted as a point estimate of the probability of failure.

In the baseline method, the failure and non-failure signatures are used as references for each group. In contrast, in the clustering techniques FCM and FSOM, each vehicle receives a degree of membership for each cluster, representing how strongly it fits within that group. For analytical purposes, when the clusters are not well defined, a minimum membership threshold is set in order to classify a vehicle as an effective member of a cluster. Vehicles with membership degrees below this threshold are considered to be diffusely associated and are used to construct the test dataset.

The threshold is selected empirically and visually, varying with the application and data behavior, to balance cluster representativeness with classification quality. Allocating diffusely associated data to the test set allows assessment of the methods' ability to identify failed vehicles in samples with less-defined patterns. However, this may affect both false-positive rates and model coverage.

4 Results

4.1 Baseline

The baseline results for both applications are presented below. This analysis serves as the reference point for assessing the performance improvements achieved through the clustering methods (FCM and FSOM).

4.1.1 Application A1

The results for the baseline method are presented in Table 3. It can be observed that the cosine metric achieved the best performance for the *P@50* (0.34) and *P@100* (0.19) metrics, whereas the Euclidean metric obtained the highest value for *P@10* (0.50). The Manhattan metric showed the lowest performance across all three cutoffs.

Additionally, Table 4 presents the proportion of failed vehicles above three different probability thresholds of failure. The number of vehicles is not an integer because active thresholds (>0.6 and >0.7) identified only failed vehicles, resulting in a failure rate of 1, albeit with a very small number of cases. With a threshold greater than 0.5, the Manhattan metric also achieved 100% failure identification, but again with a very limited number of vehicles. The cosine metric identified a larger absolute number of failures, though at a lower failure rate (0.03), due to a large number of false positives (729.20).

Table 3. Application A1 - P@N for failure identification using the **baseline** method with different distance metrics.

Metric	P@10	P@50	P@100
Euclidean	**0.50**	0.11	0.05
Cosine	0.48	**0.34**	**0.19**
Manhattan	0.32	0.08	0.05

4.1.2 Application B1

The baseline results for Application B1 are shown in Table 5 and indicate that the Cosine metric achieves a significantly superior *P@10* of 0.72, surpassing the other metrics. It also stands out in *P@50* (0.18) and @100 (0.09), demonstrating greater effectiveness in identifying similar vehicles that actually experienced failures.

Table 6 shows that from the threshold >0.6, all metrics can select only failed vehicles, with failure rates equal to or greater than 0.94. Even so, for the threshold >0.5, the Manhattan metric had a high failure rate (0.82), although with a smaller absolute number of failed vehicles identified (5.4). The Cosine metric, on the other hand, identified a higher number of failures (9.0), but with a lower failure rate (0.04), suggesting that while comprehensive, this metric may include many false positives at lower thresholds.

4.2 Clustering with FCM and FSOM

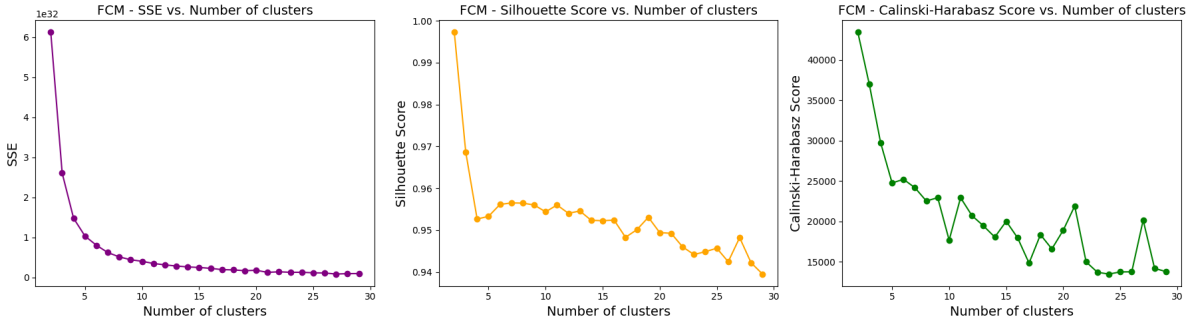
For clustering, the first step was to determine the optimal number of clusters. Figure 5a presents the evaluation metrics (SSE, Silhouette Score, and Calinski–Harabasz Index) for Application A1, suggesting that choosing $k = 3$ or $k = 4$ clusters is appropriate. The results for $k = 3$ clusters in FCM and $k = 4$ clusters in FSOM are not detailed in this article, as they revealed limitations arising from cluster-size imbalance. Therefore, only the results for FCM with $k = 4$ and FSOM with $k = 3$ are discussed in this section.

To Application B1, the number of clusters was again determined based on the SSE, Silhouette Score, and Calinski-Harabasz Index. Figure 5b indicates that both $k = 3$ and $k = 4$ are appropriate choices to segment the data.

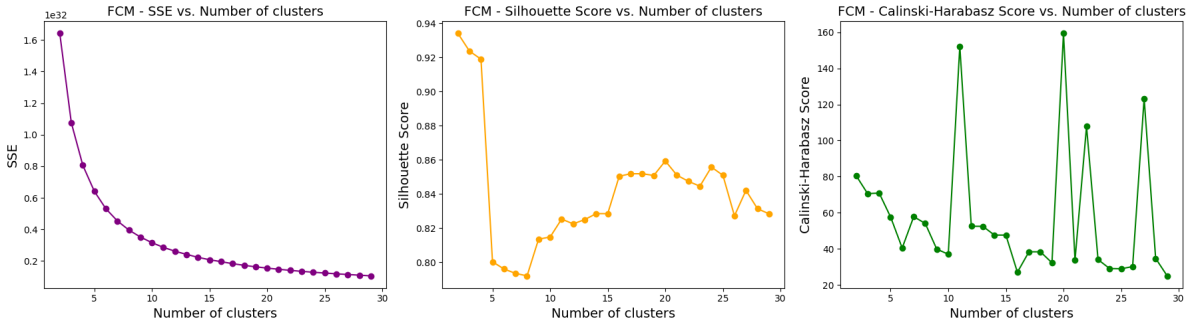
Despite some cases showing high silhouette values (e.g., FCM with $k = 3$ achieved 0.9249), FCM clustering did not yield clearly representative clusters of failed or non-failed vehicles, even when minimum membership thresholds were applied. In general, failed vehicles were scattered across groups or mixed with non-failed ones, making it impossible to construct consistent signatures based on the clusters. For FSOM with $k = 3$, results were even less satisfactory: the Silhouette was very low (0.0819), and a large number of vehicles were excluded after applying minimum membership thresholds, indicating low confidence in the generated clusters.

Table 4. Application A1 - Performance of the **baseline** method under different failure probability thresholds for each distance metric, showing failed and non-failed vehicles.

Threshold	Cosine			Manhattan			Euclidean		
	Failed	Non-failed	Failure Rate	Failed	Non-failed	Failure Rate	Failed	Non-failed	Failure Rate
> 0.5	22.00	729.20	0.03	5.40	0.00	1.00	8.80	144.20	0.06
> 0.6	4.00	0.00	1.00	2.60	0.00	1.00	3.40	0.00	1.00
> 0.7	2.60	0.00	1.00	1.20	0.00	1.00	2.20	0.00	1.00



(a) Application A1 - Evaluation metrics (SSE, Silhouette Score, Calinski–Harabasz Index) used to determine the number of clusters.



(b) Application B1 - Evaluation metrics (SSE, Silhouette Score, Calinski–Harabasz Index) used to determine the number of clusters.

Figure 5. Comparison of clustering evaluation metrics for Applications A1 and B1.

Table 5. Application B1 - P@N for identifying failed vehicles using the **baseline** method with different distance metrics

Distance Metric	P@10	P@50	P@100
Euclidean	0.40	0.08	0.08
Cosine	**0.72**	**0.18**	**0.09**
Manhattan	0.24	0.06	0.04

4.2.1 Application A1

The FCM clustering with $k = 4$ resulted in a high silhouette value (0.9527), indicating good cluster separation. However, as shown in Table 7, the groups were highly unbalanced: Most of the vehicles, defective or not, were assigned to group 4. This indicates that despite a good silhouette score, the partition was initially ineffective for signature generation. However, applying a membership threshold of > 0.55 based on the distributions shown in Figure 6a, potential signatures were identified:

- Cluster 2 as indicative of failed vehicles, with 6 failed and only 1 non-failed;
- Clusters 1 and 3 as indicative of non-failed vehicles, with 5 vehicles;
- Clusters 4 and -1 (the latter including diffusely associated vehicles) were used as the test set, given their heterogeneous composition and/or low membership in other clusters.

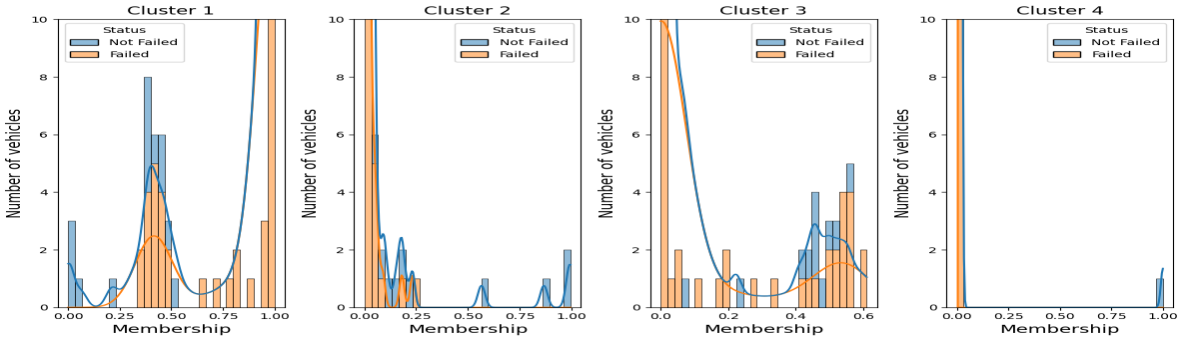
Table 8 shows the *P@N* values for failure recovery in the test dataset using different distance metrics. The Euclidean distance achieved the best result for *P@10* (1.00) but was less effective for larger lists. The cosine distance produced the most stable and highest performance for *P@50* (0.84) and *P@100* (0.79), while the Manhattan distance consistently underperformed. In general, FCM results significantly outperformed the baseline across all precision metrics, demonstrating the effectiveness of clustering with a minimum membership threshold for defining failure signatures.

Table 9 presents the performance of clustering across different failure probability thresholds. The cosine metric achieved the highest coverage at a threshold of > 0.5 (107 failed vehicles identified), although with a higher false-positive rate. At higher thresholds (> 0.6 and > 0.7), all metrics reached perfect precision but with reduced coverage. Compared to the baseline, this clustering showed slightly better performance, particularly for the Manhattan distance at the > 0.5 threshold.

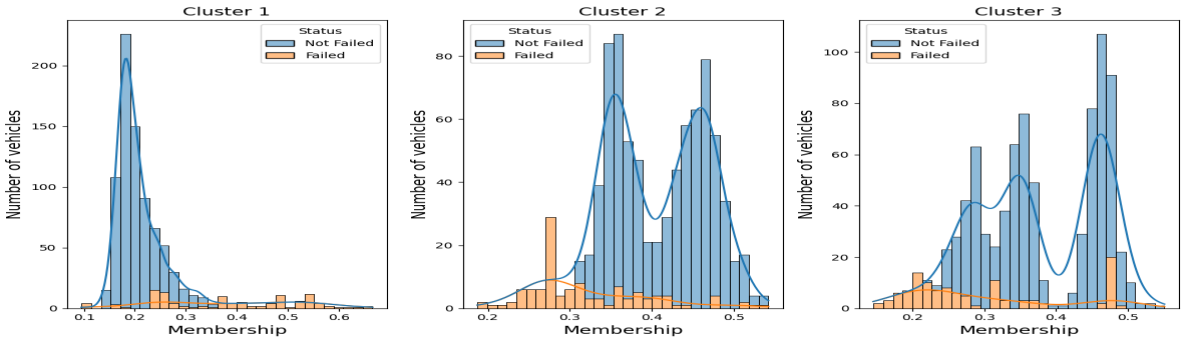
The FSOM clustering with $k = 3$ yielded a low silhouette value (0.0967), indicating weak separation between groups. Cluster 3 emerged as a clear signature of failed vehicles, comprising 48 of them, while clusters 1 and 2 remained highly heterogeneous, as shown in Table 10. Consequently, a mini-

Table 6. Application B1 - Baseline method performance at different failure probability thresholds for each distance metric, showing the number of failed and non-failed vehicles.

Threshold	Euclidean			Manhattan			Cosine		
	Failed	Non-failed	Failure Rate	Failed	Non-failed	Failure Rate	Failed	Non-failed	Failure Rate
> 0.5	4.40	41.60	0.10	5.40	1.20	0.82	9.00	197.60	0.04
> 0.6	3.00	0.00	1.00	3.40	0.20	0.94	3.00	0.20	0.94
> 0.7	1.60	0.00	1.00	2.20	0.00	1.00	2.40	0.00	1.00



(a) Application A1 - Membership distribution of vehicles using FCM ($k = 4$).



(b) Application A1 - Membership distribution of vehicles using FSOM ($k = 3$).

Figure 6. Membership distributions of vehicles in Application A1 using fuzzy clustering approaches.

Table 7. Application A1 - Comparison of the number of vehicles in each FCM cluster ($k = 4$) using (i) maximum membership and (ii) membership > 0.55. Cluster -1 includes vehicles below the threshold.

Cluster	Initial Assignment		With Membership > 0.55	
	Failed	Non-failed	Failed	Non-failed
-1			11	8
1	0	4	0	4
2	17	7	6	1
3	0	1	0	1
4	97	173	97	721

Table 8. Application A1 - *P@N* for failure identification using FCM ($k = 4$) with different distance metrics.

Distance Metric	P@10	P@50	P@100
Euclidean	**1.00**	0.46	0.26
Cosine	0.90	**0.84**	**0.79**
Manhattan	0.70	0.18	0.13

mum membership threshold of 0.55 was again applied (based on Figure 6b). As a result, clusters 1 and 2 concentrated non-failed vehicles and were considered representative of that group. Vehicles with diffuse membership (cluster -1) were used as the test dataset.

The *P@N* results for different distance metrics (Euclidean, Cosine, and Manhattan) are shown in Table 11. The Manhattan distance achieved the best precision for *P@10* (1.00), whereas the Cosine distance yielded the best results for *P@50* (0.74) and *P@100* (0.61). Although these results were again far superior to the baseline, they were slightly lower than the FCM performance for larger lists (@50 and @100). As in the FCM case, the Cosine metric remained the best choice for *P@50* and *P@100*, while the Manhattan metric stood out at P@10, unlike FCM, where the Euclidean metric performed best at that level.

The analysis by probability threshold (Table 12) again shows that increasing the threshold improves precision, reaching 100% for some metrics at higher thresholds, albeit at the expense of reduced coverage. The results were strong in all thresholds and exceeded those of FCM, with the Manhattan metric particularly notable at the 0.6 and 0.7 thresholds, identifying 25 failed vehicles with only five false positives and 15 failed vehicles with no false positives, respectively.

Table 9. Application A1 - FCM ($k = 4$) performance under different failure probability thresholds for each distance metric, showing failed and non-failed vehicles.

Threshold	Cosine			Manhattan			Euclidean		
	Failed	Non-failed	Failure Rate	Failed	Non-failed	Failure Rate	Failed	Non-failed	Failure Rate
> 0.5	107	726	0.13	11	0	1.00	21	5	0.81
> 0.6	12	0	1.00	2	0	1.00	3	0	1.00
> 0.7	5	0	1.00	1	0	1.00	2	0	1.00

Table 10. Application A1 - Distribution of vehicles by cluster in FSOM ($k = 3$) with membership threshold.

Cluster	Initial Assignment		With Membership > 0.55	
	Failed	Non-failed	Failed	Non-failed
-1			91	731
1	23	323	0	2
2	43	412	0	2
3	48	0	23	0

Table 11. Application A1 - *P@N* for FSOM ($k = 3$) with different distance metrics.

Metric	P@10	P@50	P@100
Euclidean	0.90	0.42	0.24
Cosine	0.80	**0.74**	**0.61**
Manhattan	**1.00**	0.42	0.25

For Application A1, the results demonstrate that both clustering approaches (FCM with $k = 4$ and FSOM with $k = 3$) significantly outperformed the baseline in identifying vehicles with higher failure probability. FCM showed superior performance in *P@N* metrics, particularly with the Cosine and Euclidean distances, whereas FSOM performed best for smaller lists (N = 10). Furthermore, FSOM provided more stable results across thresholds, especially with the Manhattan metric at thresholds 0.6 and 0.7, which achieved high precision with a low incidence of false positives. These findings reinforce the effectiveness of fuzzy clustering in defining failure signatures and suggest that decision thresholds and the desired size of the retrieval list can guide the choice between methods.

4.2.2 Application B1

Applying the FSOM algorithm with $k = 4$ clusters yielded a Silhouette score of 0.0646, indicating low separation among the formed groups. Figure 7 shows the distribution of vehicle membership degrees across clusters. Table 13 compares the distribution of vehicles among clusters using (i) the highest membership association and (ii) a minimum threshold of 0.40. Despite the low Silhouette, it was possible to identify two distinct clusters: cluster 1, composed exclusively of failed vehicles (even after the threshold was applied), and cluster 3, composed solely of non-failed vehicles. The remaining groups (2 and 4), as well as vehicles in cluster 1, were used as a test base to evaluate the ability to identify failures by similarity.

5 Discussion

The main findings from Applications A1 and B1 can be summarized as follows:

1. **Baseline Limitations:** The baseline method generally struggled to distinguish between failed and non-failed vehicles. Across both applications, *P@N* values remained low, except for Application B1 at *P@10* using the Cosine distance (0.72). At various failure probability thresholds, the baseline often identified few true failed vehicles despite high failure rates, indicating limited discriminative power and a tendency to produce false positives at lower thresholds.
2. **Impact of Fuzzy Clustering:** Applying Fuzzy C-Means (FCM) and Fuzzy Self-Organizing Map (FSOM) algorithms significantly improved the identification of failure and non-failure signatures by using membership thresholds to filter ambiguous cases. For A1, FCM with $k = 4$ produced high *P@N* values, particularly with Cosine and Euclidean distances, while FSOM with $k = 3$ effectively isolated high-risk vehicles. In B1, FSOM with $k = 4$ identified clusters exclusively composed of failed or non-failed vehicles, enabling more precise detection than the baseline. Both methods benefited from fuzzy membership functions to reduce uncertainty.
3. **Cluster Quality vs. Predictive Performance:** Silhouette scores did not always correlate with predictive performance. FSOM often produced low silhouette values (e.g., 0.0967 in A1, 0.0646 in B1), yet it successfully distinguished between failed and non-failed vehicles. In contrast, FCM in A1 achieved a high silhouette (0.9527) but showed cluster imbalance, with most vehicles concentrated in a single cluster. This demonstrates that practical effectiveness in failure identification depends on membership interpretation rather than solely on traditional clustering metrics.
4. **Role of Distance Metrics:** Distance metric selection directly affected performance. Cosine distance consistently produced the highest *P@N* for larger retrieval lists, effectively capturing vehicle behavior patterns. Euclidean distance performed very well for smaller lists. In contrast, Manhattan distance proved useful for high-confidence detection at low thresholds, although it was generally less effective for ranking-based metrics.
5. **Robustness and Practical Implications:** FSOM demonstrated robustness across both applications, effectively capturing non-linear and overlapping patterns even with low silhouette values. FCM offered structured, interpretable clusters, particularly valuable in A1. These complementary strengths suggest that the clustering choice and distance metric should be aligned with operational objectives, such as prioritizing early detection over minimizing false positives.
6. **Impact of Data Imbalance:** The strong class imbalance

Table 12. Application A1 - Performance of FSOM ($k = 3$) under different probability thresholds.

Threshold	Cosine			Manhattan			Euclidean		
	Failed	Non-failed	Failure Rate	Failed	Non-failed	Failure Rate	Failed	Non-failed	Failure Rate
> 0.5	83	728	0.10	34	19	0.64	19	7	0.73
> 0.6	12	3	0.80	25	5	0.83	7	0	1.00
> 0.7	8	1	0.89	15	0	1.00	2	0	1.00

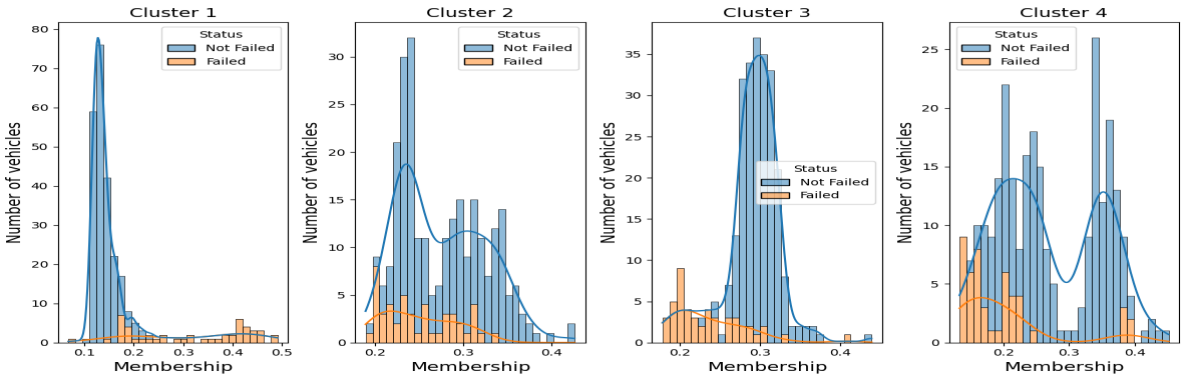


Figure 7. Application B1 - Vehicle membership distribution with FSOM ($k = 4$)

Table 13. Application B1 - Comparison of the number of vehicles in each FSOM cluster ($k = 4$) using (i) the highest membership and (ii) membership > 0.40. Cluster -1 includes vehicles below the threshold.

Cluster	Initial Assignment		With Membership > 0.40	
	Failed	Non-failed	Failed	Non-failed
-1			23	210
1	27	0	19	0
2	15	8	1	1
3	1	118	0	3
4	5	94	5	6

Table 14. Application B1 - P@N for identifying failed vehicles using FSOM ($k = 4$) with different distance metrics

Metric	P@10	P@50	P@100
Euclidean	0.90	0.26	0.16
Cosine	0.90	**0.42**	**0.27**
Manhattan	0.50	0.24	0.20

present in both applications directly influences performance interpretation. Since failed vehicles represent a small fraction of the fleet, even modest improvements in P@N correspond to meaningful gains in practical failure detection. The clustering-based approaches demonstrated the ability to concentrate failed vehicles in smaller groups despite the imbalance, indicating that the extracted behavioral features capture relevant degradation signals beyond what would be expected by random selection.

Although labeled data were available, supervised learning approaches were not evaluated in this study by design. The primary objective was to investigate whether unsupervised fuzzy clustering methods could uncover latent behavioral patterns and failure signatures without explicitly modeling the failure label. This choice was motivated by practical fleet scenarios where failure annotations may be incomplete, delayed, or noisy, particularly in large-scale telemetry systems. In addition, clustering-based methods offer greater interpretability in exploratory settings.

In addition, the feature engineering strategy adopted in this study relied on TSFRESH to transform time-series data into a static tabular representation. While this approach improves interpretability and computational efficiency, it may not fully capture sequential dependencies or temporal degradation dynamics. Deep learning architectures, such as LSTM networks or 1D convolutional neural networks, could directly model the raw temporal deltas and potentially learn hierarchical representations of failure progression. Exploring such models represents a promising direction for future research, particularly to assess if end-to-end temporal learning can further improve predictive performance.

The combination of fuzzy clustering and appropriate distance metrics significantly improved the identification of vehicles at risk of failure compared to baseline, highlighting these techniques for predictive maintenance of fleet operations.

6 Conclusion

This study evaluated data-driven strategies to proactively identify vehicles at risk of failure in Brazil's road freight sector, focusing on telemetry data from Euro 6 heavy trucks. Two approaches were compared: baseline similarity metrics and fuzzy clustering methods (FCM and FSOM). Cluster-based approaches consistently outperformed the baseline. Both *P@N* and the number of failed vehicles detected at probability thresholds increased. In A1, FCM achieved the highest precision for small and large retrieval lists, while FSOM provided robust identification of high-risk vehicles. In B1, FSOM successfully isolated clusters composed entirely of failed or non-failed vehicles, enabling more reliable failure detection. The results also suggest that operational behavior variables have stronger predictive power than individual component measurements, underscoring the importance of usage patterns in predicting failures.

Future work should consider density-based methods such

as DBSCAN or HDBSCAN, which are promising alternatives to handle non-convex structures and noisy vehicles. Supervised learning models should also be investigated to directly predict failure probabilities and extend the methodology to other engine components. Further investigation is necessary to evaluate the economic advantage of the proposed clustering approach, considering acceptable probability thresholds to replace eventual failed components. Finally, evaluating real-time deployment and scalability across large fleets would help validate the framework's practical applicability and operational benefits.

Declarations

Author's Contribution

JV, RL, HL, and TS contributed to the conception of this article. JV performed the experiments. RL, HL, and TS contributed to writing, review, and supervision. JV is the primary contributor and writer of this article. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests related to the research, authorship, and/or publication of this article.

Acknowledgements

This research was partially supported by the SocialNet project (process 2023/00148-0 of the São Paulo Research Foundation - FAPESP), by the National Council for Scientific and Technological Development - CNPq (processes 314603/2023-9, 441444/2023-7, and 444724/2024-9). This research is also part of the INCT ICoNIoT, funded by CNPq (proc. 405940/2022-0) and CAPES (Finance Code 88887.954253/2024-00).

References

Amruthnath, N. and Gupta, T. (2018). A research study on unsupervised machine learning algorithms for early fault detection in predictive maintenance. In *5th Int. Conf. on Industrial Eng. and Applications (ICIEA)*, pages 355–361. DOI: 10.1109/IEA.2018.8387124.

Azzaoui, H., Manssouri, I., and Elkihel, B. (2019). Methylcyclohexane continuous distillation column fault detection using stationary wavelet transform & fuzzy c-means. *Materials Today: Proceedings*, 13:597–606. DOI: 10.1016/j.matpr.2019.04.018.

Cao, Q., Zanni-Merk, C., Samet, A., Reich, C., de Bertrand de Beuvron, F., Beckmann, A., and Giannetti, C. (2022). KSPMI: A knowledge-based system for predictive maintenance in industry 4.0. *Robotics and Computer-Integrated Manufacturing*, 74:102281. DOI: 10.1016/j.rcim.2021.102281.

da Penha Araujo, M., Campos, V. B., and Bandeira, R. A. (2013). An overview of road cargo transport in Brazil. *Int. Journal of Industrial Eng. and Management*, 4(3):151. DOI: 10.24867/IJIEM-2013-3-119.

Fulcher, B. D. and Jones, N. S. (2014). Highly comparative, feature-based time-series classification. *CoRR*, abs/1401.3531. DOI: 10.1109/tkde.2014.2316504.

Khoshkangini, R., Sheikholharam Mashhadi, P., Berck, P., Gholami Shahbandi, S., Pashami, S., Nowaczyk, S., and Niklasson, T. (2020). Early prediction of quality issues in automotive modern industry. *Information*, 11(7). DOI: 10.3390/info11070354.

Mattos, J. G., Happ, P. N., Fernandes, W., Lopes, H. C. V., Barbosa, S. D. J., Kalinowski, M., Rosa, L. S., Novello, C., Ribeiro, L. D., Ventura, P. R., Marques, M. C., Pitta, R. N., Camolesi, V. J., Costa, L. P. L., Paravidino, B. I., and Pereira, C. S. (2023). A framework for enhancing industrial soft sensor learning models. *Digital Chemical Engineering*, 8:100112. DOI: 10.1016/j.dche.2023.100112.

Principi, E., Rossetti, D., Squartini, S., and Piazza, F. (2019). Unsupervised electric motor fault detection by using deep autoencoders. *IEEE/CAA Journal of Automatica Sinica*, 6(2):441–451. DOI: 10.1109/JAS.2019.1911393.

Ravi, A., Surabhi, M., and Shah, C. (2022). Machine learning applications in predictive maintenance for vehicles: Case studies. *Int. Journal of Eng. and Computer Science*, 11:25628–25640. DOI: 10.18535/ijecs/v11i08.4707.

Samatas, G. G., Moumgiakmas, S. S., and Papakostas, G. A. (2021). Predictive maintenance - bridging artificial intelligence and IoT. In *2021 IEEE World AI IoT Congress (AIIoT)*, pages 0413–0419. DOI: 10.1109/AIIoT52608.2021.9454173.

Schlechtingen, M. and Santos, I. (2011). Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection. *Mechanical Systems and Signal Processing*, 25:1849–1875. DOI: 10.1016/j.ymssp.2010.12.007.

Seixas, L. D., Corrêa, F. C., Siqueira, H. V., Trojan, F., and Afonso, P. (2023). Vehicle industry big data analysis using clustering approaches. In *Int. Conf. on Optimization, Learning Algorithms and Applications*, pages 312–325. Springer. DOI: 10.1007/978-3-031-53036-$4_2$2.

Shaowu, S., Sheng, Z., Wanlu, J., and Zhenbao, L. (2020). Study on the health condition monitoring method of hydraulic pump based on convolutional neural network. In *2020 12th Int. Conf. on Measuring Technology and Mechatronics Automation (ICMTMA)*, pages 149–153. IEEE. DOI: 10.1109/icmtma50254.2020.00041.

Singpurwalla, N. D. and Wilson, S. P. (1998). Failure models indexed by two scales. *Advances in Applied Probability*, 30(4):1058–1072. DOI: 10.1239/aap/1035228207.

Surucu, O., Gadsden, S. A., and Yawney, J. (2023). Condition monitoring using machine learning: A review of theory, applications, and recent advances. *Expert Systems with Applications*, 221:119738. DOI: 10.1016/j.eswa.2023.119738.

Tagliatti, C., Melanda, E. A., and Martins, D. d. O. (2024). Logistics infrastructure in Brazil – an overview of the cargo transport system. *Observatorio de La Economía Latinoamericana*, 22(10):e7306. DOI: 10.55905/oelv22n10-162.

Viscenheski, J. R., Lüders, R., Lopes, H., and Silva, T. H.

(2025). Towards predictive maintenance of heavy-duty trucks exploring telemetry and warranty data. In *2025 21st Int. Conf. on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT)*, pages 586–592. DOI: 10.1109/DCOSS-IoT65416.2025.00094.

Wasserman, G. S. (1992). An application of dynamic linear models for predicting warranty claims. *Computers Industrial Engineering*, 22(1):37–47. DOI: 10.1016/0360-8352(92)90031-E.