www.nature.com/scientificdata

scientific data

Check for updates

OPEN

DATA DESCRIPTOR

SCANIA Component X dataset: a real-world multivariate time series dataset for predictive maintenance

Zahra Kharazian[1]✉, Tony Lindgren[1,2], Sindri Magnússon[1], Olof Steinert[2] & Oskar Andersson Reyna[3]

Predicting failures and maintenance time in predictive maintenance is challenging due to the scarcity of comprehensive real-world datasets, and among those available, few are of time series format. This paper introduces a real-world, multivariate time series dataset collected exclusively from a single anonymized engine component (Component X) across a fleet of SCANIA trucks. The dataset includes operational data, repair records, and specifications related to Component X while maintaining confidentiality through anonymization. It is well-suited for a range of machine learning applications, including classification, regression, survival analysis, and anomaly detection, particularly in predictive maintenance scenarios. The dataset's large population size, diverse features (in the form of histograms and numerical counters), and temporal information make it a unique resource in the field. The objective of releasing this dataset is to give a broad range of researchers the possibility of working with real-world data from an internationally well-known company and introduce a standard benchmark to the predictive maintenance field, fostering reproducible research.

Background & Summary

In an era marked by technological advancement and data-driven decision-making, the automotive industry is undergoing a transformative shift, particularly in the realm of vehicle maintenance. Predictive maintenance (PdM) has emerged as a key player in revolutionizing how we care for and optimize the performance of vehicles. This innovative methodology harnesses the power of advanced analytics, sensor technology, and machine learning to predict when vehicle components are likely to fail, allowing for timely and cost-effective maintenance interventions. PdM is paramount for critical components in trucks. This proactive approach to maintenance involves using data, analytics, and monitoring systems to estimate the components' health state.

One of the significant challenges in PdM is the shortage of public real-world datasets. The reason is that Original Equipment Manufacturers (OEMs) tend to keep the data to themselves and do not share it with anyone outside their company. Exceptions are research or development partners who can access the data when working with companies for PdM solutions. Reasons for keeping the data hidden are that it is sensitive as it contains failure frequencies, what type of sensor is available, etc. The lack of data makes researchers in the field use simulated datasets[1–4] that mimic real conditions. However, having access to real-world datasets remains essential for developing robust models for foreseeing real industrial equipment failures, as synthetic datasets usually lack the complexity of real-world data and overlook many common challenges, such as intricate relations between signals.

The newly released real-world dataset[5] from SCANIA is an exception to the typical practice of companies not releasing data publicly and holds significant potential for advancing the field of PdM. Another unique advantage inherent in this dataset is its consideration of temporal information, demonstrating gradual degradation in equipment in the form of time series readouts. This multi-variate time series dataset is likely rich in diverse information collected from more than 33'000 SCANIA trucks and can be used for various tasks in PdM, such as classification, regression, forecasting, anomaly detection, and survival analysis. The proposed dataset is introduced for the Industrial Challenge 2024 at the 22nd International Symposium on Intelligent Data Analysis (IDA) with the title of *Developing an Effective Predictive Model for Imminent Component X Failures in Heavy-Duty SCANIA Trucks* at Stockholm University, Sweden. The setting of this industrial challenge served as a nexus for

[1]Stockholm University, Department of Computer and Systems Sciences, Kista, SE-164 07, Sweden. [2]Scania CV, Strategic Product Planning and Advanced Analytics, Södertälje, SE-151 32, Sweden. [3]Scania CV, Connected Intelligence, Södertälje, SE-151 32, Sweden. ✉e-mail: zahra.kharazian@dsv.su.se

collaboration, fostering a unique synergy between academia and industry. The proposed dataset opens the door to various opportunities for advancing research and optimizing PdM methodologies.

This dataset includes operational data, truck specifications, and the repair records of an anonymized truck engine component called component X. The component's name has been anonymized for proprietary reasons. The wealth of information in the dataset empowers decision-makers to make informed choices regarding maintenance strategies, resource allocation, and fleet management. This dataset has been used in recent PdM-related research studies; for instance, Zhong *et al*.[6] implemented deep learning models such as convolutional and recurrent neural networks to detect failure patterns and maintenance cost management. Parton *et al*.[7] employed a method using graph neural networks (GNNs) on the graphs that are derived from this time series data to estimate the remaining useful life (RUL). Moreover, Carpentier *et al*.[8] applied different models in various applications like multiclass classification, regression, and survival analysis to predict component X's failure. Additionally, another study[9] used the dataset to evaluate a novel loss and error calculation method designed for survival analysis and RUL prediction. In a different application[10], it has been used for low dimensional synthetic data generation to enhance the predictive model's performance. Another study[11] highlights the effect of implementing active learning in enhancing the model's decision for RUL estimation on this dataset. In another study[12], machine learning-based survival analysis models are applied, and SHAP-based explainability is used to improve transparency in RUL prediction for component X.

Another comparable dataset also comes from SCANIA, it was concerned with the Air Pressure System (APS)[13] of trucks and was disclosed in a preceding industrial challenge 2016 at the 15th International Symposium on Intelligent Data Analysis (IDA), Stockholm University. The APS dataset is also suitable for many machine learning tasks in the PdM field and has been used in various research, such as[14–26]. Although the APS dataset is rich in the field, it does not capture the temporal information of variables.

The rest of this paper is structured as follows: The Methodology section discusses the data collection process, associated challenges, and the steps taken to preserve data privacy through anonymization. The Data Records section provides a detailed description of the dataset. Finally, the Technical Validation section introduces a cost function designed to evaluate the performance of a trained model on the dataset.

Methods

Data collection.

The proposed dataset includes three sources of information encompassing important aspects of truck information. The first part comprises operational data. The second source contains repair records obtained from workshops. The last one incorporates the specifications of trucks. Collecting such information from a fleet of trucks operating daily in different situations involves several challenges and potential errors. Below, we explain how the data is collected and also mention some of the possible errors during data collection. Finally, we explain what considerations have been made to protect the confidentiality of data for publishing.

Operational data. For collecting operational data of this dataset, trucks' onboard sensors are utilized to monitor and collect crucial parameters like real-time data on the truck's condition and performance. This data is stored inside the vehicle control units and is accessible remotely or using plugin cables in visiting workshops. One challenge in compiling operational data is losing connection with devices that collect data. One example of this error happens with Electronic Control Units (ECUs), like the engine control unit; sometimes, when the ECU software is updated, the data collection counters could be reset, i.e., start again from the beginning. This type of problem with data is handled through post-processing of the data when it has been downloaded from the truck. The post-processing does cover the majority of the possible types of corrupt data when collecting information but it might not cover all cases.

Repair records. Repair records collected from trucks include information about maintenance, repairs, and servicing performed on the vehicles. This information is usually collected from invoices and work orders, including crucial information about services rendered, replaced parts, labor costs, and other expenses incurred during the workshop visits. If the components have been changed or marked as repaired, they will be considered as failed and labeled as such in the dataset. Otherwise, it will be considered as healthy. Here, the challenge is that SCANIA can receive information only from its own network and official workshops. Therefore, we limited the population of the vehicles to those with a complete service history.

Specifications of trucks. Specification data is collected with the production system. Where they provide detailed specifications for each truck's model the company has produced, this includes information about the engine type, weight capacities, dimensions, and other technical details. One possible challenge in collecting such data is, in very rare cases, when the truck is rebuilt and no longer matches the original specification.

Some of the mentioned errors are handled or addressed using quality control measures and regular equipment maintenance. Also, collaboration between data scientists, engineers, and domain experts has been done as a crucial step for developing effective strategies to handle data collection challenges in PdM.

Preserving Data Privacy.

After collecting data, some modifications have been made to protect the confidentiality of data for publishing. Regarding temporal representation description, relative times are reported instead of the original timestamps. This is done to capture temporal patterns without revealing specific dates or time points. Repair frequencies and readout frequencies may have been modified and are not necessarily representative of actual truck usage. Moreover, variable names have been omitted for privacy and proprietary reasons, presenting a subset of all available operational and specification data. Also, the dataset comprises a random subset of vehicles visiting SCANIA workshops, and real vehicle identity numbers are not disclosed and are reported as anonymous IDs. The chosen subset is representative of SCANIA workshop visits for Component

X analysis. In addition to these considerations, some perturbations like scaling have been made on operational data and repair rates. It is worth noting that the perturbed data still maintains its utility for predictive modeling. To evaluate the robustness of the model, a sensitivity analysis is conducted, assessing how changes in perturbation levels impact the model's performance. This allowed us to observe model effectiveness variations under different perturbation conditions. Furthermore, it is validated that perturbations do not compromise the predictive model's accuracy.

The anonymization of the Component X dataset is necessary to protect proprietary and sensitive information, and without such modifications, it would not have been possible to publish this dataset. While this process may add complexity to the interpretation of certain features, it is important to note that the annonymized data retains its utility for predictive modeling. Moreover, the dataset's inherent temporal and operational patterns remain intact. These patterns further form a basis for training and evaluating predictive models. To facilitate the connection between anonymized data parts, a unique ID is associated with each vehicle across all shared files. This enables researchers to integrate information from different data parts, such as operational data and time-to-event information, ensuring that comprehensive analyses can still be conducted effectively despite anonymization. This dataset supports a wide range of applications, including predicting the remaining useful life of components, identifying fault patterns, detecting anomalies, and optimizing maintenance schedules. Additionally, the dataset can be utilized for survival analysis, classification, and regression tasks, enabling researchers to address diverse challenges in PdM and develop robust, data-driven solutions.

Data Records

The proposed dataset is publicly available and can be accessed on the Swedish National Data Service website (https://doi.org/10.5878/jvb5-d390)[5]. It is divided into three segments: training, validation, and testing. Each segment comprises multiple files in CSV (Comma-Separated Values) format, and this section provides detailed descriptions for each file to facilitate ease of use. Vehicles are randomly selected for inclusion in each segment to ensure the dataset's robustness and reliability, promoting a representative distribution across the entire dataset. This random selection process effectively allocates a specific percentage (70%, 15%, and 15%) of vehicles to the training, validation, and testing sets, respectively, allowing for comprehensive model evaluation and development.

Training set.

The training set contains three files including "*train_operational_readouts.csv*", "*train_tte.csv*", and "*train_specifications.csv*," where each is explained below.

train_operational_readouts.csv. Operational data related to the train set is collected in a file named "*train_operational_readouts.csv*". This data file comprises readouts collected at different times from a variety of features for a fleet of vehicles, yielding a multi-variate time series. These readouts, found in each row, offer unique insights into the operational status of these vehicles. Each readout encapsulates a distinct set of variables gathered between two points in time, such as t_i and t_{i+1}. In summary, it consists of 1'122'452 observations or instances from 23'550 unique vehicles and 107 columns, including *vehicle_id* and *time_step*. The column *time_step* acts as a gauge, measuring the duration in time_step that each vehicle has been utilizing Component X during its operational lifespan. Note that vehicles do not necessarily follow the same sampling frequency in the time_step column. It is also worth noting that this dataset[5] does not encompass the entire available operational data but rather represents a carefully selected subset. Experts have specifically curated this subset, handpicking data they believe is most relevant. Ultimately, 14 attributes are selected and anonymized in the operational data, offering a broad spectrum of information without divulging specifics about the nature of Component X. These variables are organized into single numerical counters and histograms where each histogram has several bins and provides a compressed representation of the signal data by grouping the sensor readings into bins. The basis for the division of the bins varies from the range of the data (maximum and minimum of the variable), the variable characteristics, the resolution of the data, experts' knowledge, and industry-specific standards. Each bin in a histogram represents certain conditions linked to the values observed within the measured features. For instance, imagine a histogram linked to the variable *distance_driven* with four bins representing *ambient temperature*. This histogram shows the distribution of distance driven, organized into bins with different temperature ranges $\{(T < -20), (-20 \leq T < 0), (0 \leq T < 20), (T > 20)\}$. Each bar of this histogram shows a temperature range, and the height of each bar represents the frequency of *distance_driven* within that temperature range. Histogram variables use the following indexing format: *variableid_binindex*. Where the "variableid" represents the ID of an anonymized variable or feature, and "binindex" shows the bin numbers. As an example, the variable with "variableid" 167 is a multi-dimensional histogram that has ten bins, "167_0", "167_1",..., and "167_9".

In summary, six out of 14 variables are organized into six histograms with variable IDs: "167", "272", "291", "158", "459", and "397," with 10, 10, 11, 10, 20, and 36 bins, respectively. Figure 1 illustrates two histogram features of variables 167 (see Fig. 1a) and 459 (see Fig. 1b) from an arbitrary vehicle at its last readout. Moreover, the eight rest of the variables named "171_0", "666_0", "427_0", "837_0", "309_0", "835_0", "370_0", "100_0"] are numerical counters. These features are mostly accumulative and are suitable for the representation of trends over time. Figure 2 visualizes these numerical counters in more detail. An example of these features for an arbitrary vehicle in this dataset[5] is depicted in Fig. 2a.

Additionally, the correlation between these non-histogram features (numerical counters) is calculated considering all readouts of vehicles and illustrated in Fig. 2b. As can be seen, all the features are positively correlated, and no negative correlation exists between them. i.e., if the value of one feature increases, the value of the other feature also tends to increase.

The distribution of missing values in *train_operational_readouts.csv* is shown in Fig. 3. In this dataset, all the readouts were collected directly from the ECUs connected to the sensors of the heavy-duty trucks. The missing

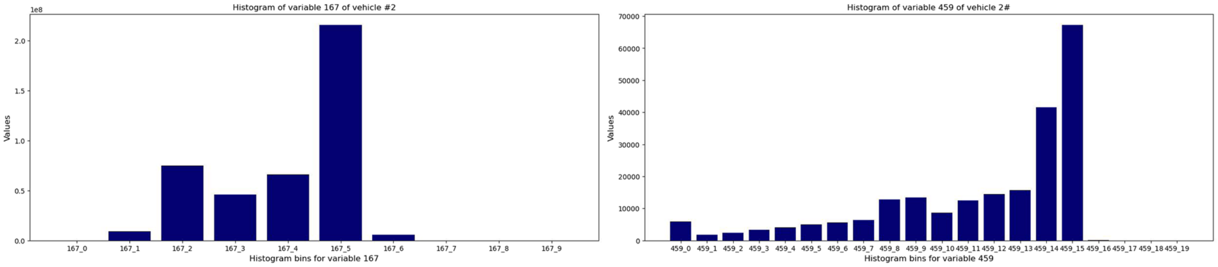

Fig. 1 Visualization of two histogram variables for vehicle number 2. On the left, (**a**) displays histogram variable 167, while on the right, (**b**) illustrates histogram variable 459.

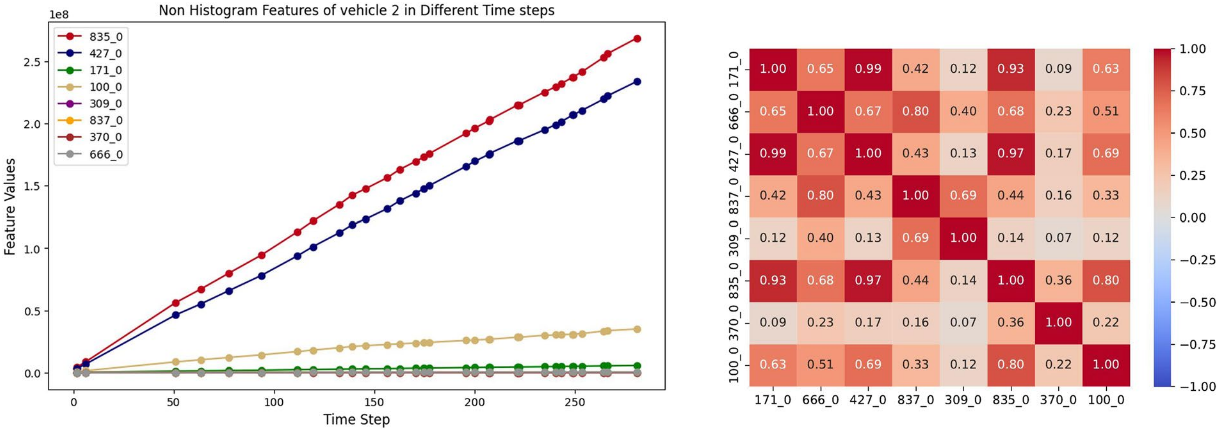

Fig. 2 Analyzing non-histogram features over time for the train set. On the left, (**a**) demonstrates non-histogram variables for vehicle number 2 over time, while on the right, (**b**) shows the correlation of these features.

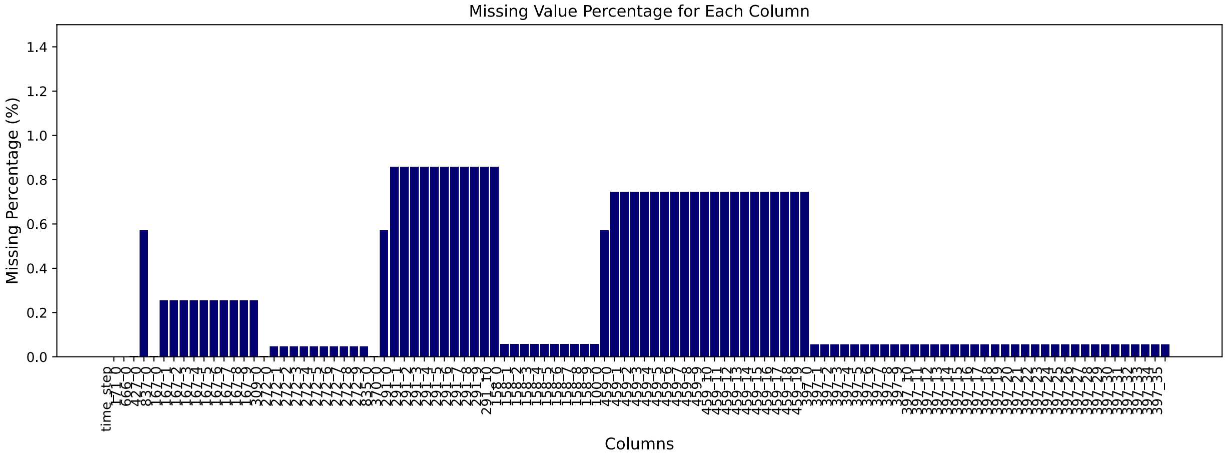

Fig. 3 The missing value percentage in train_operational_readouts.csv file is less than 1% per feature column. Note that the y-axis shows the percentage of missing and is limited to 1.5 for better visualization.

values occurred for various reasons, such as the vehicle not being equipped with the sensor in question, the ECU software not being set to log the signal, or communication issues with the ECU. However, due to the high quality of the sensors and software, the rate of missing values is low, with less than 1 percent missingness per feature/column. Given the low percentage of missing values, the dataset's integrity remains relatively intact, allowing various machine learning tasks with minimal preprocessing.

train_tte.csv. The file with the name "*train_tte.csv*" contains the repair records of Component X collected from each vehicle, indicating the time_to_event (tte), i.e., the replacement time for Component X during the

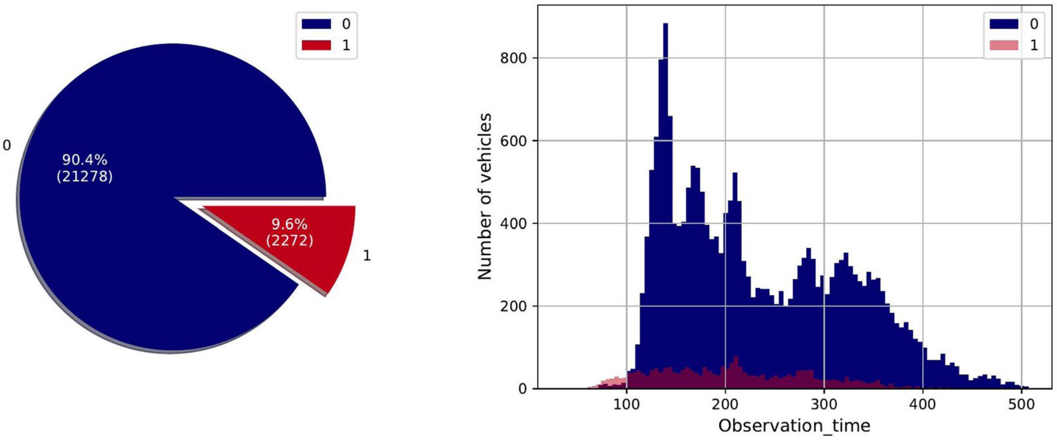

Fig. 4 Visualization of classes in the train set. On the left, (**a**) shows the repair distribution, while on the right, (**b**) illustrates the distribution of healthy and failed components with different observation times during data collection.

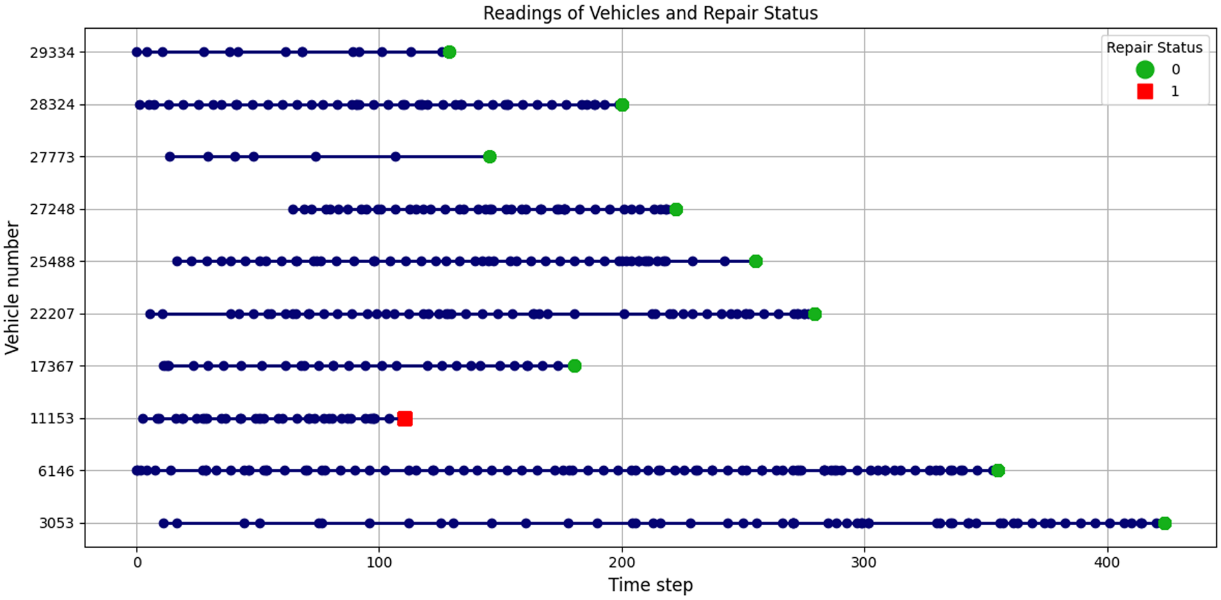

Fig. 5 History of readouts for ten random vehicles.

study period. This data file includes 23'550 number of rows and two columns: "*length_of_study_time_step*" and "*in_study_repair*," where the former indicates the number of operation time steps after Component X started working. The latter is the class label, where it's set to 1 if Component X was repaired at the time equal to its corresponding *length_of_study_time_step*, or it can take the value of zero in case no failure or repair event occurs during the first *length_of_study_time_step* of operation. It is good to mention that the "*train_tte.csv*" data is imbalanced with 21'278 occurrences of label 0 and 2'272 instances of label 1. In other words, it is skewed toward label 0. Figure 4a compares the number of healthy and repaired components in the train set in general. Moreover, Fig. 4b shows the distribution of healthy and repaired components in their corresponding observation time during the data collection, i.e., the time between the last and first readout for each vehicle.

Delving deeper into the history of readouts in the training data, Fig. 5 illustrates the time of the readouts for ten random vehicles. In this figure, blue dots represent the individual readout events for each vehicle and illustrate the monitoring frequency of events in the *train_operational_readout* data file. Besides, the final readout for each vehicle is collected from the *train_tte* data part and highlighted with colors: Green circles show components having healthy status in their last workshop visits, while red squares mark repaired components. It also can be seen that the readouts are distributed unevenly among different vehicles. In addition to this information, there are no missing values (shown by NaN) in this data file.

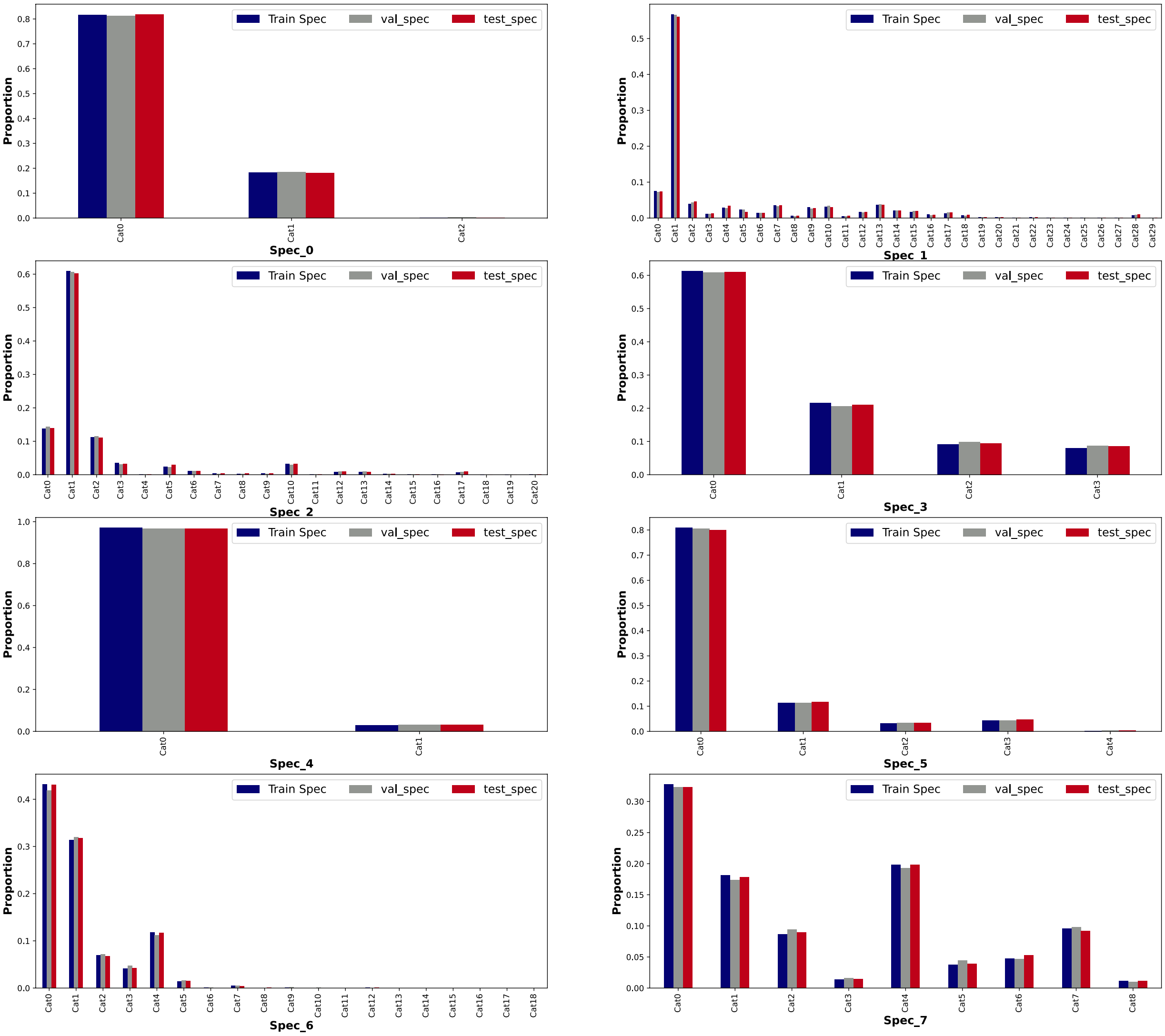

Fig. 6 Frequency analysis of specification (normalized) for training, validation, and test set.

train_specifications.csv. The last file in the training set is called "*train_specifications.csv*," which contains information about the specifications of the vehicles, such as their engine type and wheel configuration. In total, there are 23'550 observations and eight categorical features, with no missing values for all vehicles.

The features in *train_specifications.csv* are anonymized, each can take categories in Cat0, Cat1, …, Cat28. Figure 6 illustrates the normalized frequency distribution of this categorical data for different data parts (train, validation, and test), allowing for comparison of how each category is represented across the three datasets. Each subplot represents a different specification (e.i., Spec_0 to Spec_7). In each subplot, the x-axis shows the categorical values that each specification feature can take, and the y-axis depicts the normalized proportion of each category across the data parts with different colors: dark blue, grey, and red for train, validation, and test set, respectively. Since the proportions are similar across data parts, this indicates that the category distributions are relatively consistent.

Validation set. The validation set consists of three files called "*validation_labels.csv*", "*validation_operational_readouts.csv*", and "*validation_specification.csv*".

validation_operational_readouts.csv. In general, the *validation_operational_readouts.csv* has the same description as the *train_operational_readouts.csv* except for the fact that in *validation_operational_readouts.csv*, the operational data is incomplete. Only a subset of the whole observations of each vehicle is provided, and it extends only up to a randomly selected readout. As a result, it lacks details about the entire lifespan of a vehicle. This is done to simulate the usage of a prediction model in a realistic scenario when it only has information about a vehicle up until the present time. Figure 7 illustrates an example of a hypothetical health indicator or degradation model of component X installed on a vehicle in the validation set. Green dots are recorded readouts from the start of its operation, and the yellow star represents the last simulated readout for this vehicle, which

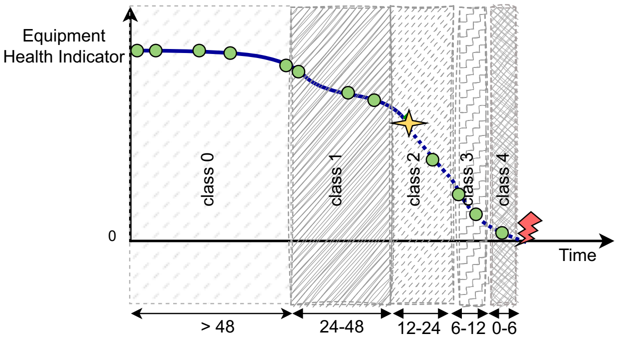

Fig. 7 Class interpretation in the validation and test data.

is randomly selected among all possible readouts. This means that we only have information up to that readout, and the rest of the information is not given.

Overall, it includes 196'227 observations/rows showing the number of instances from 5046 vehicles and includes 107 columns. Moreover, similar to *train_operational_readouts.csv*, it contains a minimal missing value (less than one percent for each feature).

validation_labels.csv. The *validation_labels.csv* file has 5046 rows, which is equal to the number of vehicles contributed to the operational data of the validation set. It includes a column named *class_label*, corresponding to the class for the last readout of each vehicle. As mentioned in subsection *validation_operational_readouts.csv*, the last readout for the validation set is selected randomly among all readouts for each vehicle. The temporal placement of this final simulated readout is categorized into five classes denoted by 0, 1, 2, 3, 4 where they are related to readouts within a time window of: (more than 48), (48 to 24), (24 to 12), (12 to 6), and (6 to 0) time_step before the failure, respectively. These classes show the time windows in which the last readouts for each vehicle are randomly selected. For instance, in Fig. 7, the last simulated readout is given in the time window of class 2.

This data set is also imbalanced and is skewed toward class 0, i.e., 4910 samples belong to class 0, while 76, 30, 16, and 14 samples belong to classes 4, 3, 1, and 2, respectively.

For better visualization of the validation data, two techniques for dimensionality reduction, called Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE), are performed on the last readout from each vehicle in the validation set. Figure 8a and b show the dataset when its dimension is reduced into two, using PCA and t-SNE techniques, respectively. Vehicles belonging to each class are shown in different colors. As can be seen, the five classes of the dataset are scrambled in the two-dimensional space both for PCA and t-SNE. This demonstrates the complexity of the problem when the features are projected into two dimensions.

validation_specification.csv. The file *validation_specification.csv* has a similar structure as *train_specifications.csv* with no missing values, except the data is collected from 5046 vehicles in the validation set. Details of categorical values of each feature are shown in Fig. 6. Compared to the bars related to the *train_specifications*, the grey bars related to the *validation_specification*, follows the same normalized distribution, where shows the consistency in categories across both data parts.

Testing set.

The test set contains three files "*test_operational_readouts.csv*", "*test_specifications.csv*", and "*test_labels.csv* which are explained in detail in the following sections."

test_operational_readouts.csv. Similar to *validation_operational_readouts.csv*, for the *test_operational_readouts.csv*, the last readouts of vehicles are randomly selected from the larger sequences observed during the study period (see Fig. 7) to emulates real-life situations. In summary, this file contains 198'140 number of readouts from 14 variables (107 columns) gathered from 5045 unique vehicles. Similar to *validation_operational_readouts.csv* and *train_operational_readouts.csv*, the percentage of missing values in this file is less than one percent for each feature.

test_labels.csv. This data file includes the class label for 5045 vehicles in the test set that their last readouts are randomly selected in five classes of 0, 1, 2, 3, and 4. Like the *validation_labels*, this data file is also imbalanced and is skewed toward class 0, which contains 4903 samples. In comparison, classes 1, 2, 3, and 4 include 26, 15, 41, and 60 samples.

test_specifications.csv. Specification information of 5045 test vehicles is collected in *the test_specifications.csv* file containing eight categorical features with values varying between Cat0, Cat1,..., and Cat28, with no missing values. Figure 6 illustrates the categories of each feature in the *(*_specifications.csv* files, including for the test set. Compared to the bars related to the train_specifications and validation_specifications, the red bars representing the test_specification exhibit the same normalized distribution, indicating the consistency in category distribution across all three data parts.

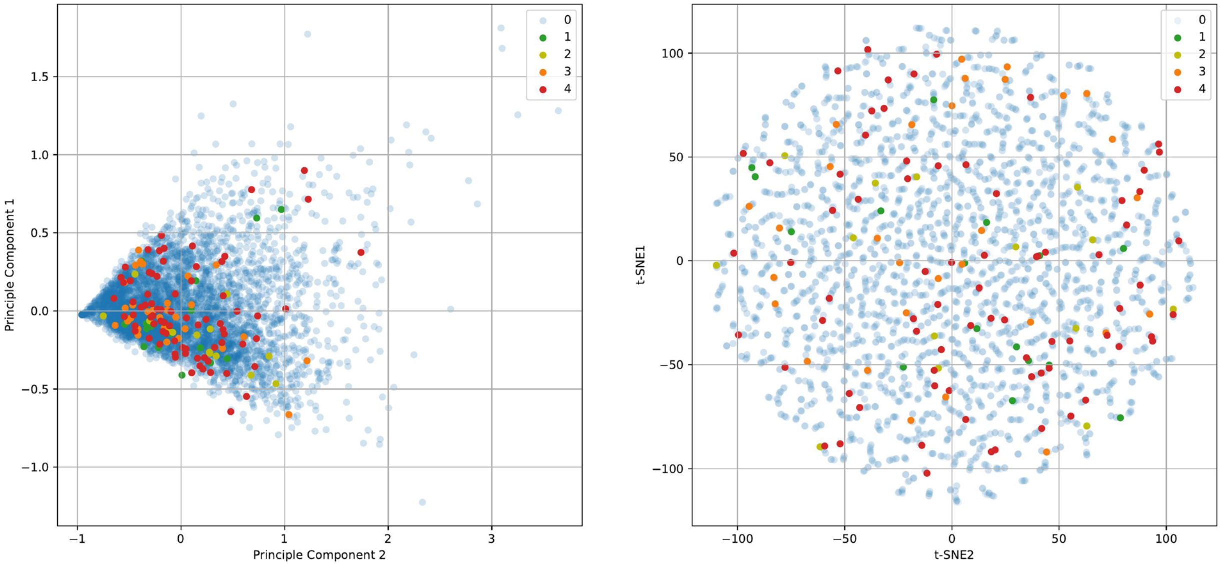

Fig. 8 Visualization of validation data using dimensionality reduction techniques. Classes are shown in different colors. On the left, (**a**) presents the PCA analysis of the validation set, while on the right, (**b**) displays the t-SNE analysis.

	Predicted: 0	Predicted: 1	Predicted: 2	Predicted: 3	Predicted: 4
Actual: 0		Cost_0_1=7	Cost_0_2=8	Cost_0_3=9	Cost_0_4=10
Actual: 1	Cost_1_0=200		Cost_1_2=7	Cost_1_3=8	Cost_1_4=9
Actual: 2	Cost_2_0=300	Cost_2_1=200		Cost_2_3=7	Cost_2_4=8
Actual: 3	Cost_3_0=400	Cost_3_1=300	Cost_3_2=200		Cost_3_4=7
Actual: 4	Cost_4_0=500	Cost_4_1=400	Cost_4_2=300	Cost_4_3=200	

Table 1. Table of prediction cost.

Technical Validation

OEMs usually hesitate to share their device operational data for many reasons, such as GDPR compliance, confidentiality agreements (NDAs), ownership of data, device failure rates, etc. As a result, many works on data-driven PdM are limited to private datasets[27–32]. This prevents researchers who worked with real data comparing their methods performances with each other. The newly shared dataset[5] by SCANIA can serve as a benchmark in the field, making it easier for future researchers to compare their methods and generate reproducible results.

The proposed dataset[5] is highly conducive to various machine learning tasks like regression, anomaly detection, survival analysis, and classification, in the PdM field. This merit arises from two supreme characteristics: being a real-world dataset collected from actual trucks and exhibiting a multi-variate time series structure. In regression tasks, this dataset[5] is valuable for estimating components' remaining useful life or predicting the time until the next repair. Leveraging historical and time-to-event data, regression models can provide insightful estimates. Survival analysis is a powerful technique that predicts the survival function and the probability of an event happening at a specific time while considering censored data in the training phase. The blue vehicle populations in Fig. 4 are referred to as censored data. The temporal structure of the dataset is crucial for the effective application of survival analysis. Furthermore, for classification tasks, using time series data, the model can classify whether a vehicle is going to fail within a specific time window or not.

To evaluate the performance of the aforementioned models, a cost function is suggested by experts of the company, which is defined by the sum of the different "Cost_*n*_*m*" multiplied by the number of instances, resulting in a summarized cost (Total_cost).

$$Total\ \_cost\ =\ Cost\_n\_m\ \times\ No\_\ instances \tag{1}$$

Equation (1) calculates the total cost function, where *n* shows the actual class, and m shows the predicted class, while $n, m \in \{0,1,2,3,4\}$. In general, when $n < m$, the Cost_*n*_*m* indicates a cost for a false positive error, and if $n > m$, it indicates a cost for a false negative error. It should be emphasized that the cost of false negative prediction is much higher than that of false positive prediction. Table 1 demonstrates the different values of Cost_*n*_*m* according to different values of actual and predicted class prediction. This table formulates a spectrum of the cost of failure from a catastrophic one (i.e., unexpected failure by the road) to a non-necessary cost of changing the component. The values of different costs in Table 1 are defined by the experts of the company. In this table, the term "Cost_0_{1,2,3,4}" denotes the cost incurred when a mechanic conducts an unnecessary check at a workshop. On the other hand, "Cost_{1,2,3,4}_*m*" represents the cost associated with the possibility of missing a faulty truck or triggering an alarm too late. This delay could lead to a breakdown or necessitate costly adjustments to the customer's transportation plan.

Code availability

All the readings are collected from sensors installed on heavy-duty trucks. The data is then transferred to the company's database when the trucks visit the workshops or through telemetry. The raw readings were then converted to the CSV (Comma-Separated Values) format. Subsequently, the dataset is curated, anonymized, and validated using custom codes in Python.

Received: 21 February 2024; Accepted: 11 March 2025;
Published online: 24 March 2025

References

1. Turbofan Engine Degradation Simulation Data Set. NASA Prognostics Data Repository, NASA Ames Research Center, Moffett Field, CA https://www.nasa.gov/content/diagnostics-prognostics (2008).
2. Arias Chao, M., Kulkarni, C., Goebel, K. & Fink, O. Aircraft engine run-to-failure dataset under real flight conditions for prognostics and diagnostics. *Data* **6**, 5 (2021).
3. Voronov, S., Frisk, E. & Krysander, M. Data-driven battery lifetime prediction and confidence estimation for heavy-duty trucks. *IEEE Transactions on Reliability* **67**, 623–639 (2018).
4. Safdari, A., Frisk, E., Holmer, O. & Krysander, M. Synthetic generation of streamed and snapshot data for predictive maintenance. *IFAC-PapersOnLine* **58**, 270–275 (2024).
5. Lindgren, T., Steinert, O., Andersson Reyna, O., Kharazian, Z. & Magnússon, S. SCANIA Component X Dataset: A Real-World Multivariate Time Series Dataset for Predictive Maintenance, https://doi.org/10.5878/jvb5-d390 (2024).
6. Zhong, J. & Wang, Z. Implementing deep learning models for imminent component x failures prediction in heavy-duty scania trucks. In *International Symposium on Intelligent Data Analysis*, 268–276 (Springer, 2024).
7. Parton, M., Fois, A., Vegliò, M., Metta, C. & Gregnanin, M. Predicting the failure of component x in the scania dataset with graph neural networks. In *International Symposium on Intelligent Data Analysis*, 251–259 (Springer, 2024).
8. Carpentier, L., De Temmerman, A. & Verbeke, M. Towards contextual, cost-efficient predictive maintenance in heavy-duty trucks. In *International Symposium on Intelligent Data Analysis*, 260–267 (Springer, 2024).
9. Rahat, M. & Kharazian, Z. Survloss: A new survival loss function for neural networks to process censored data. In *PHM Society European Conference*, vol. 8, 7–7 (2024).
10. Lindgren, T. & Steinert, O. Low dimensional synthetic data generation for improving data driven prognostic models. In *2022 IEEE International Conference on Prognostics and Health Management (ICPHM)*, 173–182 (IEEE, 2022).
11. Kharazian, Z., Lindgren, T., Magnússon, S. & Boström, H. Copal: Conformal prediction for active learning with application to remaining useful life estimation in predictive maintenance. *Proceedings of Machine Learning Research* **230**, 1–23 (2024).
12. Kargar-Sharif-Abad, M., Kharazian, Z., Miliou, I. & Lindgren, T. SHAP-Driven Explainability in Survival Analysis for Predictive Maintenance Applications. In ECAI: EUROPEAN CONFERENCE ON ARTIFICIAL INTELLIGENCE, HAII5. 0: Embracing Human-Aware AI in Industry 5.0 (2024).
13. APS Failure at Scania Trucks. UCI Machine Learning Repository, https://doi.org/10.24432/C51S51 (2017).
14. Huang, Z., Wu, Y., Tempini, N., Lin, H. & Yin, H. An energy-efficient and trustworthy unsupervised anomaly detection framework (eatu) for iiot. *ACM Transactions on Sensor Networks* **18**, 1–18 (2022).
15. Abidi, M. H., Umer, U., Mohammed, M. K., Aboudaif, M. K. & Alkhalefah, H. Automated maintenance data classification using recurrent neural network: Enhancement by spotted hyena-based whale optimization. *Mathematics* **8**, 2008 (2020).
16. Selvi, K. T., Praveena, N., Pratheksha, K., Ragunanthan, S. & Thamilselvan, R. Air pressure system failure prediction and classification in scania trucks using machine learning. In *2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, 220–227 (IEEE, 2022).
17. Lokesh, Y., Nikhil, K. S. S., Kumar, E. V. & Mohan, B. G. K. Truck aps failure detection using machine learning. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 307–310 (IEEE, 2020).
18. Ranasinghe, G. D., Lindgren, T., Girolami, M. & Parlikad, A. K. A methodology for prognostics under the conditions of limited failure data availability. *IEEE Access* **7**, 183996–184007 (2019).
19. Oh, E. & Lee, H. Quantum mechanics-based missing value estimation framework for industrial data. *Expert Systems with Applications* **236**, 121385 (2024).
20. Ke, Q., Siłka, J., Wieczorek, M., Bai, Z. & Woźniak, M. Deep neural network heuristic hierarchization for cooperative intelligent transportation fleet management. *IEEE Transactions on Intelligent Transportation Systems* **23**, 16752–16762 (2022).
21. Sun, K., Magnússon, S., Steinert, O. & Lindgren, T. Robust contrastive learning and multi-shot voting for high-dimensional multivariate data-driven prognostics. In *2023 IEEE International Conference on Prognostics and Health Management (ICPHM)*, 53–60 (IEEE, 2023).
22. Akarte, M. M. & Hemachandra, N. Predictive maintenance of air pressure system using boosting trees: A machine learning approach. In *ORSI* (2018).
23. Rafsunjani, S., Safa, R. S., Al Imran, A., Rahim, M. S. & Nandi, D. An empirical comparison of missing value imputation techniques on aps failure prediction. *International Journal of Information Technology and Computer Science* **2**, 21–29 (2019).
24. Syed, M. N., Hassan, M. R., Ahmad, I., Hassan, M. M. & De Albuquerque, V. H. C. A novel linear classifier for class imbalance data arising in failure-prone air pressure systems. *IEEE Access* **9**, 4211–4222 (2020).
25. Taghandiki, K. & DallakehNejad, M. Minimizing the repair cost of the air pressure system of scania trucks using a deep learning algorithm. *TechRxiv* (2023).
26. Beikmohammadi, A. *et al*. A cost-sensitive transformer model for prognostics under highly imbalanced industrial data. *arXiv preprint arXiv:2401.08115* (2024).
27. Moat, G. & Coleman, S. Survival analysis and predictive maintenance models for non-sensored assets in facilities management. In *2021 IEEE international conference on big data (Big Data)*, 4026–4034 (IEEE, 2021).
28. Rahat, M., Kharazian, Z., Mashhadi, P. S., Rögnvaldsson, T. & Choudhury, S. Bridging the gap: A comparative analysis of regressive remaining useful life prediction and survival analysis methods for predictive maintenance. In *PHM Society Asia-Pacific Conference*, vol. 4 (2023).
29. Hoffmann, M. A. & Lasch, R. Roadmap for a successful implementation of a predictive maintenance strategy. *Smart and Sustainable Supply Chain and Logistics–Trends, Challenges, Methods and Best Practices: Volume 1* 423–439 (2020).
30. Rahat, M., Pashami, S., Nowaczyk, S. & Kharazian, Z. 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).

31. Bouabdallaoui, Y., Lafhaj, Z., Yim, P., Ducoulombier, L. & Bennadji, B. Predictive maintenance in building facilities: A machine learning-based approach. Sensors **21**, 1044 (2021).
32. Rahat, M. *et al*. Domain adaptation in predicting turbocharger failures using vehicle's sensor measurements. In *Phm society european conference*, vol. 7, 432–439 (2022).

Acknowledgements

This work has been funded by Scania CV AB and the Vinnova Program for Strategic Vehicle Research and Innovation (FFI).

Author contributions

Z.K. conceived data analysis, visualization, writing - original draft, writing - review & editing. T.L. conducted experimental design, writing - review & editing and holding the chair position for the industrial challenge associated with the definition of this dataset. S.M. helped with writing - review & editing. O.S. conducted experimental design, writing - review & editing. O.A. contributed to experimental design and data collection. All authors reviewed the manuscript.

Funding

Open access funding provided by Stockholm University.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Z.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.