CEUR-WS.org/Vol-2568/paper10.pdf

Data-Driven Powertrain Component Aging Prediction Using In-Vehicle Signals

Andreas Udo Sass[1], Enes Esatbeyoglu[1], and Till Iwwerks[1]

Volkswagen AG, Brieffach 17772, 38436 Wolfsburg, Germany

Abstract. Predictive maintenance has become an important tool to avoid unplanned downtime of modern vehicles. The exchanged data between Electronic Control Units (ECU) is simultaneously increasing with the functionality. A large number of in-vehicle signals are provided and facilitate the monitoring of physical component aging processes. In this work, we generated a training dataset and observed aging of a selected powertrain component.

First, we preprocessed in-vehicle signals to generate a time equidistant signal database. Furthermore, the signals were segmented in various time periods and subsequently aggregated to statistical features. Second, the signal associated aging information were synchronized to an equal time frame. We investigated several signals preselection approaches to predict an aging-value for the powertrain component with machine learning methods. These approaches differ in the count of selected in-vehicle signals for the aging-value prediction.

Our results show that in-vehicle signals can be used to predict powertrain component aging. The quality of estimation differs with respect to the selected regression methods. In this work we present an approach to narrow down the prediction quality of different preselection approaches for the estimation of a powertrain component aging.

Keywords: Predictive maintenance · feature extraction · signal preselection · time series · machine learning.

1 Introduction

Various amounts of time-resolved data is recorded in the life-cycle of vehicles. This data is transmitted from Electronic Control Units (ECU) via an Control Area Network (CAN) bus of the vehicle. The complexity of modern vehicles grows rapidly. Many components in the vehicle communicate with each other. A reliable diagnosis of an potential aging of a component is complex.

Predictive maintenance in a commercial mobility context let the customer know the current status of his vehicle(s). There is no extensive definition of predictive maintenance. Hence, it is defined in various ways according to its use in literature [1]. On the one hand, predictive maintenance estimates a possible system or component failure. On the other hand, the Remaining Useful Life (RUL) can be predicted as a health management.

Copyright (c) 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

In our paper we implemented a health management or Remaining Useful Life (RUL) prediction. The RUL prediction with provided raw monitoring data is presented in [2] or with additional sensors in [3–5]. Instead of estimating a RUL the condition-based maintenance (CBM) gives recommendations concerning maintenance decisions [6]. We estimate a degree of aging of an Exhaust Gas Recirculation (EGR) cooling system. In sense of predictive maintenance the RUL can be derived from a given aging degree. Different vehicles are equipped with CAN-Loggers to record the in-vehicle signals. These information are the results of the communication from different ECUs and contain sensors readings, actuators readings and internal parameters of control models. The transferred information on the CAN bus is not equidistant. Because of the arbitration, various messages have different priorities. The requested CAN information is related to the actual driving state. The real-time performance of the CAN bus is analyzed by comparison between the time-triggered and event-triggered protocol in [7].

The paper is structured as follows. Section 2 describes various data-driven diagnostic applications in the literature. Moreover, the physical EGR cooling system aging effect is introduced and data preprocessing workflow is explained in detail. Section 3 presents the result for modelling the aging-value of the given powertrain component. We compare the quality of different signal preselection approaches and different regression methods. Section 4 concludes with a discussion and gives an outlook of future work.

2 Background

In this section, we provide background information to predict the aging degree of a selected powertrain component by using different regression methods. We use machine learning algorithms to create a model of the aging degree. The training set's target value (ground truth) is given by observing the fouling of the Exhaust Gas Recirculation (EGR) cooler components in certain intervals in a workshop. Data-driven diagnostics is applied in the automotive domain to analyze vehicle components and support manufacturer and vehicle driver decision making. For example, the On-board Diagnostics (OBD) system monitors fault diagnosis of vehicle components and notifies the driver regarding the possible malfunction of vehicle component. This is initially designed to keep the vehicle emissions within statutory thresholds [8].

Besides the monitoring of emission limits with OBD systems, other authors use machine learning algorithms to apply data-driven diagnostics in the automotive domain. Machine learning algorithms within the context of regression problems generate a functional dependency between input data and target values, without explicit being programmed for it. Training data based on in-vehicle signal logging is used to fit a machine learning model. The learned model is subsequently applied to make predictions on new datasets.

The data-driven diagnostics for component aging prediction presented by several papers differ in the source and amount of data used for training the model. On one hand, models are trained with data from special sensors. For that purpose some authors use vibration sensors or acoustic emissions [4, 5, 9, 10]. On the other hand, some works use a dataset without adding special sensors. To reduce the whole input data and for optimizing the models results, some authors select a subset of the most relevant information [11, 12]. Instead of using in-vehicle data or data from additional sensors, some authors [13] use maintenance meta-information of a vehicle fleet for forecasting.

Different machine learning algorithms are used for fault diagnosis. The input data is labeled concerning a fault state or normal state. Support Vector Machines (SVM) are used to predict the fault state [14]. The authors of [15] use Bayesian Networks for the same purpose. Some authors use simple Neural Networks (NN) to predict failures within the given input data [16–18]. Wolf et al. combine several Neural Networks to predict the preignition of high-pressure turbocharged petrol engines [19]. Instead of an offline calculation of the models, a cloud can calculate complex machine learning algorithms of the transmitted data [20]. Because of a limited transfer rate between the car and the cloud, only a reduced information amount is transmitted [16]. Some authors apply predictive maintenance methods in the industrial sector [21]. Automotive components like bearings, gears and shafts are analyzed regarding their common features [22, 23]. Instead of detecting fault states, some authors search healthy representatives in the data. These representatives are used for monitoring, diagnostics and prognostics. This is exemplified for an automotive braking system in [24].

2.1 EGR Component

The specific vehicle component investigated in this work is analyzed with the help of in-vehicle date from several vehicle prototypes with diesel engines. To keep the legal pollutant emission limit a recirculation of exhaust gas is applied in these prototypes.

At an ideal combustion of sulfur-containing diesel fuel carbon dioxide (CO2), water (H2O) and sulfur dioxide (SO2) are released as resulting products. The air conditions are fluctuating strong locally. Due to a non-ideal combustion, nitrogen oxides (NOx), carbon monoxide (CO), hydrocarbons (HC) and particulate matter (PM) are created. The EGR valve controls the recirculating exhaust gas of the engine back into the intake tract.

Nitrogen oxides (NOx) emissions can be reduced by increasing the EGR rate. In addition to that the implemented system uses a cooler to decrease the gas temperature. Therefore the peak temperature can be reduced. Another benefit is the possibility to control and reduce NOx emissions from diesel engines by decreasing the combustion temperature [25, 26]. However, a higher EGR rate promotes an increased fouling in EGR coolers [27, 28]. This fouling affects the cooler performance negatively. A built-in bypass switch controls the exhaust gas flow to be cooled, if necessary. This can influence the hydrocarbon and carbon monoxide emissions positively [29].

The EGR cooler fouling consists of HC and PM deposits. Due to these deposits, the flow resistance rises. To reduce the quantity of deposits an additional catalyst in the EGR line can be implemented [30]. The EGR cooler fouling is a complex process and also dependent on the engine operating state [31].

2.2 Data Description

The following explanations describe the two data sources used for further analysis and model prediction. The first data source includes the in-vehicle signals from several prototypes. The prototypes itself were used for road trails especially designed to collect data related to the EGR aging process. The second data source is provided through data collected during workshop visits. This includes the measured degree of the EGR cooling system aging that was measured in certain intervals. The two data sources were originally not time synchronized. The in-vehicle signals from the first source are recorded from the internal vehicle network (CAN bus) of each prototype. The data is recorded in form of time series containing series of tuples with measurement value and associated timestamps. As described in section 1, the recorded in-vehicle signals are independent and have not the same timely resolution. Owing to this, the signals were synchronized by means of their related time-stamps. We use time-series data with time resolution of 100 ms in time-equidistant form. For some binary signals (e.g. binary status signals) it is necessary not to interpolate between the signal values. For this reason the time stamps of the signals are changed to the given 100 ms grid. In order to train a machine learning model, all data vectors have to be the same size. Thus, the signal length of all analyzed signals has to be the same. At the recording start the engine could be idle and in-vehicle signals are not transmitted yet. After turning off the engine, the recording device is still working for some time and collecting all remaining information on the CAN bus. To secure the same value length over several engine starts, all in-vehicle signals have to be cut according a trigger signal. The trigger signal should be accessible on every prototype over the whole analysis period (see fig. 1). The second data source provides the aging degree of each prototype's EGR cooling system. The unevenly spaced time series (target value information) were interpolated to the

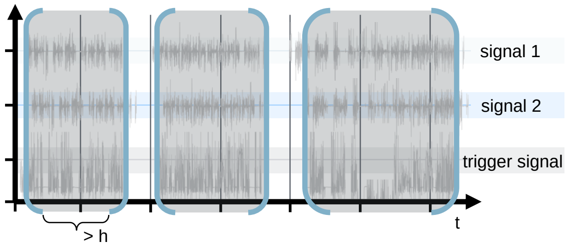

Fig. 1. Example for cut-off signals based on a trigger signal.

same 100 ms grid as mentioned above. But in this case, the aging effect occurs not only while measuring, but during the whole vehicle usage. At the given prototypes the fouling of EGR cooling system increase due to their usage.
The aging degree is defined as mass flow ratio $\dot{m}_r$ between the air mass flows of different states. While measuring the mass flow ratio the driving state has to be constant, especially the engine speed and torque. For this reason the measurements are operated in certain intervals workshops. The air mass flow is quantified while the EGR valve is in a closed and opened state. The equidistant and 100 ms synchronized aging degree is used as target value (ground truth) for further analysis.

$$\dot{m}_r = \frac{\dot{m_{f_o}}}{\dot{m_{f_c}}} \tag{1}$$

$\dot{m}_r$: mass flow ratio, degree for EGR cooler aging
$\dot{m_{f_o}}$: mass flow, EGR valve open
$\dot{m_{f_c}}$: mass flow, EGR valve closed

3 Results

In this section, we present results for the aging degree prediction of the prototypes. We choose different preselection approaches to determine prediction relevant signals (physical and data-driven approach). The modeled aging degrees were compared to the measured ones. The analyzed signals were segmented according different time periods (10 minutes to 15 hours). We used the root-mean-square error (RMSE) as performance measure to evaluate the models performance. The RMSE indicates how well the average modeled aging estimation of the respective model is compared to the target value. The smaller the RMSE, the better the model aging estimation.

Data Preselection The in-vehicle signals were preselected regarding three different approaches. The first one is the physical approach. In Section 2.2 the aging degree is given as a function of the mass flow. In a previous work we show results concerning selecting the right in-vehicle signals to monitor this powertrain component aging in dynamic working conditions for each prototype. Furthermore an optimal time period for the data aggregation is given in [32]. We use the same selection of signals described in this work as physical approach. The signals are: EGR valve position, EGR mass flow and the information about an active EGR cooling.
The extended physical approach includes the preselection of relevant signals according the theoretical information given in Section 2.1. The in-vehicle signal preselection for the extended physical approach is supplemented by signals of working conditions and internal cooling temperatures. In total we preselected ten signals.

The data-driven approach selects all in-vehicle signals, which are present on all prototypes. As a reason of having different engine configurations and recording loggers the signal selection distinguishes. After the signal intersection there are still more than 130 signals valid for analysis the aging degree.

Data Segmentation In addition to the preselection, the data is segmented in various time periods from 10 minutes till 15 hours. In each time period an associated target value is calculated concerning the given input data. Although the measured value is not as highly time-resolved as the time period, we interpolate the target value for the given time period as long as the aging process between two measurements is continuous. Thus, for each time period the ground truth is provided as aging value and we calculate a modeled aging value with different approaches by the given aggregated input data.

Data Aggregation In order to aggregate the data, we evaluate different statistical features for each of the signals used in the given segmentation. The selected statistical features are: arithmetical mean, 25th and 75th percentile and the standard deviation of the values in each time period. The statistical features of each data segmentation were used to train a multiple linear regression, Bayesian linear regression and Random Forest regression model.

Implementation In order to estimate the aging-value, we implement a multiple linear regression, a Bayesian ridge regression and a Random Forest regression [33]. A linear regression describes the relation between the dependent variable Y and the matrix of predictors X. The multiple linear regression returns the vector β of coefficients to be estimated, Y is dependent on the predictors X [34]. The multiple linear regression determines β as a estimation by minimizing the error vector ε with the least squares method. Y is defined as:

$$Y = X\beta + \varepsilon \tag{2}$$

In contrast to the multiple linear regression, the Bayes regression assumes that the errors e are independent and normally distributed random variables. The response Y is estimated not as a single point, but as a result of a probability distribution.

$$\varepsilon \sim N(0, \sigma^2) \tag{3}$$

A Random Forest regression is related to the decision tree regression. The decision tree is learning the rules of if-else sequences. At the end of these trees, numerical predictions are calculated for those leaves. The Random Forest regression combine several tree decision under a certain kind of randomization to prevent overfitting their training set.

The Figure 2 shows the estimated degree of aging for the three different regression methods for a selected prototype. For each sample consisting of statistical features an estimated aging degree is calculated using the mentioned regression methods. For the training of the regression model we use the datasets of three vehicles. The dataset of a fourth and fifth vehicle is used for the inter-vehicle validation of our approach. The RMSE is calculated for each regression methods and each data preselection approach. The physical and the extended physical preselection approach deliver a quite precise degree of aging for the selected component. The data driven approach is not as good as the other approaches, especially while using the multiple linear and Bayes regression. Instead of presenting the estimated aging degree, we compare different preselection approaches and various data segmentations in Figure 3. As described above, the model is trained with the dataset of three prototypes. The model is validated with the dataset of two separate prototypes. The Random Forest regression has a quite similar result for both physical approaches for the two selected prototypes. The Random Forest regression has better predictions than the other two for the given input.

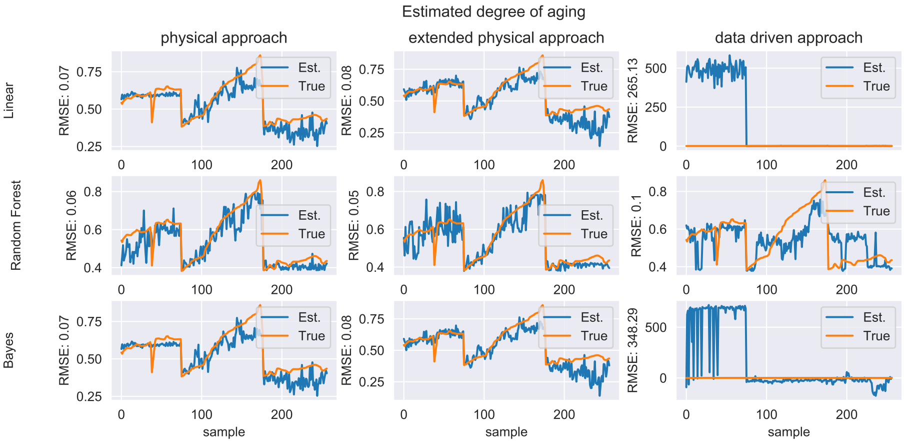

Fig. 2. Overview of different in-vehicle data preselection approaches with variation of target estimators based on data segmentation of 10 hours. Used methods are: multiple linear regression, Bayesian linear regression and Random Forest regression.

Furthermore, the physical selection approach provides a lower RMSE for the multiple linear and Bayesian regression. The RMSE of the extended physical approach is slightly worse than that. All visible regression methods have in common, that the estimation quality gets significant higher for a data segmentation above 2 hours. The data-driven preselection approach is not plotted, because the RMSE of that approach is not as precise as the other approaches.

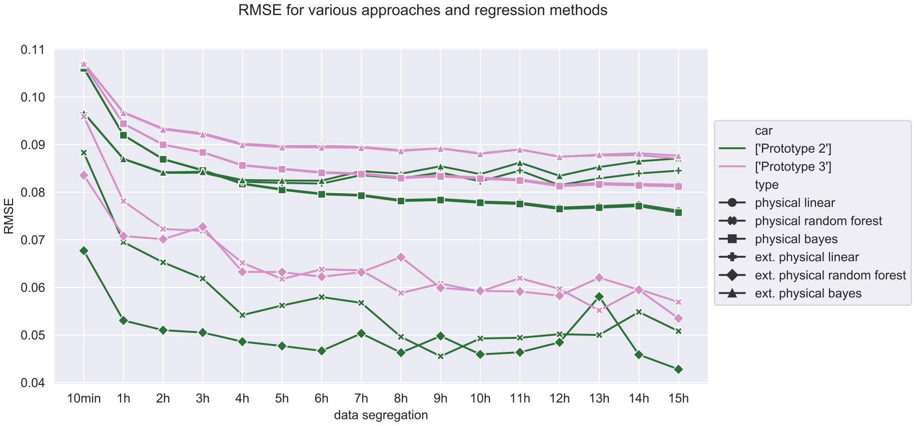

Fig. 3. RMSE overview regarding different data segmentation and the two physical preselection approaches. Used methods are: multiple linear regression, Bayesian linear regression and Random Forest regression.

4 Conclusion and Future Work

In this work we analyzed dynamic in-vehicle signals regarding the aging process of the EGR cooling system. Our approach enables the possibility to estimate the aging-value of this structural component using different preselected datasets.
First, the data from dynamic in-vehicle signals are preprocessed to generate a time equidistant dataset. Thus, it is possible to predict the aging degree by using unevenly spaced time series. Moreover, several vehicles with a various set of signals can be utilized to train the model. For each vehicle, the degree of aging is observed in certain time intervals and a training dataset is generated. This aging degree is interpolated to the same time grid as the in-vehicle signals. We provided different data preselection approaches in order to enable the aging estimation on a reduced dataset. Afterwards, we segmented the data in several time periods. The aggregated statistical features of the preselected signals were used to train the models. Figure 2 and 3 show that a prediction of a degree of aging for a selected component is feasible by using the right amount of signals. It can be noted that the quality of the prediction evidently depends on the data segmentation and on the preselected signals used for model training. As shown in Figure 3, the quality of prediction is dependent on the selected data preselection approach. The Random Forest regression finds relevant features also in the bigger dataset for the given vehicle.
The preselection approach of in-vehicle signals can be extended in consideration of each signal's relevance for the physical aging process. In this context, the relevance of each signal can be weighted concerning the physical context and used for the model training. The aim is to train the model with a subset of data to get an optimal performance.

Our goal was to deliver component aging indicators for the usage of predictive maintenance. Predictive maintenance comprises not only the prediction of component failures. In this context, predictive maintenance tries to define the degree of aging by using in-vehicle signals. In particular, for predicting this aging degree relevant signals have to be identified. Besides the determination of the aging degree, predictive maintenance is understood as notification system, in which part of the vehicle a possible aging process can be detected. For that a list of relevant in-vehicle signals should be generated.
In the future, the waveform characteristics of various component aging processes can be stored. With the help of this characteristics the relevant in-vehicle signals can be detected. Furthermore, the results of the future approach are the list of in-vehicle signals, which are identified to be relevant for a given aging process. This signal list indicates, which component aging occurs in the given dataset.

References

1. Prytz, R.: Machine Learning Methods for Vehicle Predictive Maintenance Using Off-Board and on-Board Data. PhD thesis, Halmstad University, Halmstad (2014)
2. Tobon-Mejia, D., Medjaher, K., Zerhouni, N.: CNC machine tool's wear diagnostic and prognostic by using dynamic Bayesian networks. Mechanical Systems and Signal Processing **28** (April 2012) 167–182
3. Goyal, D., Pabla, B.: Condition based maintenance of machine tools—A review. CIRP Journal of Manufacturing Science and Technology **10** (August 2015) 24–35
4. Teti, R., Jemielniak, K., O'Donnell, G., Dornfeld, D.: Advanced monitoring of machining operations. CIRP Annals **59**(2) (2010) 717–739
5. Bediaga, I., Mendizabal, X., Arnaiz, A., Munoa, J.: Ball bearing damage detection using traditional signal processing algorithms. IEEE Instrumentation & Measurement Magazine **16**(2) (April 2013) 20–25
6. Jardine, A.K., Lin, D., Banjevic, D.: A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mechanical Systems and Signal Processing **20**(7) (October 2006) 1483–1510
7. Albert, A., Strasser, R., Trächtler, A.: Migration from CAN to TTCAN for a Distributed Control System. In: Proceedings of 9th International CAN in Automation Conference. Volume 5. (2003) 9–16
8. Posada, F., Bandivadekar, A.: Global overview of on-board diagnostic (OBD) systems for heavy-duty vehicles. Int. Counc. Clean Transp. http://www.theicct.org/sites/default/files/publications/ICCT_Overview_OBD-HDVs_20150209. pdf (2015)
9. Caesarendra, W., Kosasih, B., Tieu, A.K., Zhu, H., Moodie, C.A., Zhu, Q.: Acoustic emission-based condition monitoring methods: Review and application for low speed slew bearing. Mechanical Systems and Signal Processing **72-73** (May 2016) 134–159
10. Papacharalampopoulos, A., Stavropoulos, P., Doukas, C., Foteinopoulos, P., Chryssolouris, G.: Acoustic Emission Signal Through Turning Tools: A Computational Study. Procedia CIRP **8** (2013) 426–431
11. Xu, J., Wang, Y., Xu, L.: PHM-Oriented Sensor Optimization Selection Based on Multiobjective Model for Aircraft Engines. IEEE Sensors Journal **15**(9) (September 2015) 4836–4844

12. Subrahmanya, N., Shin, Y.C., Meckl, P.H.: A Bayesian machine learning method for sensor selection and fusion with application to on-board fault diagnostics. Mechanical Systems and Signal Processing **24**(1) (January 2010) 182–192
13. Gardner, J., Koutra, D., Mroueh, J., Pang, V., Farahi, A., Krassenstein, S., Webb, J.: Driving with Data: Modeling and Forecasting Vehicle Fleet Maintenance in Detroit. arXiv:1710.06839 [cs] (October 2017)
14. Wang, Y., Ma, Q., Zhu, Q., Liu, X., Zhao, L.: An intelligent approach for engine fault diagnosis based on Hilbert–Huang transform and support vector machine. Applied Acoustics **75** (January 2014) 1–9
15. Guo, W., Zhu, Z., Hou, Y.: A novel fault diagnosis for vehicles based on time-varied Bayesian network modeling. In: 2011 Chinese Control and Decision Conference (CCDC), Mianyang, China, IEEE (May 2011) 1504–1508
16. Guo, H., Crossman, J.A., Murphey, Y.L., Coleman, M.: Automotive signal diagnostics using wavelets and machine learning. IEEE transactions on vehicular technology **49**(5) (2000) 1650–1662
17. Moosavi, S.S., N'Diaye, A., Djerdir, A., Ait-Amirat, Y., Arab Khaburi, D.: Artificial neural network-based fault diagnosis in the AC–DC converter of the power supply of series hybrid electric vehicle. IET Electrical Systems in Transportation **6**(2) (June 2016) 96–106
18. Zhang, Z., He, H., Zhou, N.: A neural network-based method with data preprocess for fault diagnosis of drive system in battery electric vehicles. In: 2017 Chinese Automation Congress (CAC), Jinan, IEEE (October 2017) 4128–4133
19. Wolf, P., Mrowca, A., Nguyen, T.T., Baker, B., Gunnemann, S.: Pre-ignition Detection Using Deep Neural Networks: A Step Towards Data-driven Automotive Diagnostics. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, IEEE (November 2018) 176–183
20. Shafi, U., Safi, A., Shahid, A.R., Ziauddin, S., Saleem, M.Q.: Vehicle Remote Health Monitoring and Prognostic Maintenance System. Journal of Advanced Transportation **2018** (2018) 1–10
21. Lee, J., Wu, F., Zhao, W., Ghaffari, M., Liao, L., Siegel, D.: Prognostics and health management design for rotary machinery systems—Reviews, methodology and applications. Mechanical Systems and Signal Processing **42**(1-2) (January 2014) 314–334
22. Carino, J.A., Delgado-Prieto, M., Iglesias, J.A., Sanchis, A., Zurita, D., Millan, M., Ortega Redondo, J.A., Romero-Troncoso, R.: Fault Detection and Identification Methodology Under an Incremental Learning Framework Applied to Industrial Machinery. IEEE Access **6** (2018) 49755–49766
23. Carino, J.A., Delgado-Prieto, M., Zurita, D., Millan, M., Ortega Redondo, J.A., Romero-Troncoso, R.: Enhanced Industrial Machinery Condition Monitoring Methodology Based on Novelty Detection and Multi-Modal Analysis. IEEE Access **4** (2016) 7594–7604
24. Magargle, R., Johnson, L., Mandloi, P., Davoudabadi, P., Kesarkar, O., Krishnaswamy, S., Batteh, J., Pitchaikani, A.: A Simulation-Based Digital Twin for Model-Driven Health Monitoring and Predictive Maintenance of an Automotive Braking System. In: The 12th International Modelica Conference, Prague, Czech Republic, May 15-17, 2017. (July 2017) 35–46
25. Ladommatos, N., Balian, R., Horrocks, R., Cooper, L.: The Effect of Exhaust Gas Recirculation on Combustion and NOx Emissions in a High-Speed Direct-injection Diesel Engine. In: International Congress & Exposition. (February 1996)

26. Zelenka, P., Aufinger, H., Reczek, W., Cartellieri, W.: Cooled EGR - A Key Technology for Future Efficient HD Diesels. In: International Congress & Exposition. (February 1998)
27. Hoard, J., Abarham, M., Styles, D., Giuliano, J.M., Sluder, C.S., Storey, J.M.E.: Diesel EGR Cooler Fouling. SAE International Journal of Engines **1**(1) (October 2008) 1234–1250
28. Bravo, Y., Moreno, F., Longo, O.: Improved Characterization of Fouling in Cooled EGR Systems. In: SAE World Congress & Exhibition. (April 2007)
29. Eitel, J., Kramer, W., Lutz, R.: Abgasrückführung: Neue Abgaskühler reduzieren Emissionen von Dieselmotoren. ATZ - Automobiltechnische Zeitschrift **105**(9) (September 2003) 856–859
30. Styles, D., Curtis, E., Ramesh, N., Hoard, J., Assanis, D., Abarham, M., Sluder, S., Storey, J., Lance, M.: EGR cooler fouling–visualization of deposition and removal mechanisms. In: Proceedings of the Directions in Engine-Efficiency and Emissions Research (DEER) Conference, Detroit, MI, October. (2011) 3–6
31. Bravo, Y., Larrosa, C., Arnal, C., Gargiulo, V.: Untersuchung der Ablagerungsbildung bei AGR-Kühlern. MTZ-Motortechnische Zeitschrift **76**(5) (2015) 36–41
32. Sass, A.U., Esatbeyoglu, E., Fischer, T.: Monitoring of Powertrain Component Aging Using In-Vehicle Signals. Diagnose in mechatronischen Fahrzeugsystemen XIII: Neue Verfahren für Test, Prüfung und Diagnose von E/E-Systemen im Kfz (2019) 14
33. Liaw, A., Wiener, M.: Classification and regression by randomForest. R news **2**(3) (2002) 18–22
34. Chatterjee, S., Hadi, A.S., et al.: Influential observations, high leverage points, and outliers in linear regression. Statistical science **1**(3) (1986) 379–393