Feature importances as a tool for root cause analysis in time-series events

Michał Kuk[1][0000−0002−6270−3938], Szymon Bobek[2][0000−0002−6350−8405], Bruno Veloso[3][0000−0001−7980−0972], Lala Rajaoarisoa[4][0000−0001−9624−5843], and Grzegorz J. Nalepa[2][0000−0002−8182−4225]

[1] AGH University of Science and Technology, Krakow, Poland
[2] Jagiellonian University, Faculty of Physics, Astronomy and Applied Computer Science, Institute of Applied Computer Science, and Jagiellonian Human-Centered AI Lab (JAHCAI), and Mark Kac Center for Complex Systems Research, ul. prof. Stanisława Łojasiewicza 11, 30-348 Kraków, Poland
[3] Faculty of Economics - University of Porto and INESC TEC, 4200-072 Porto, Portugal
[4] CERI Digital Systems IMT Nord Europe, University of Lille F-59000, France

Abstract. In an industrial setting, predicting the remaining useful life-time of equipment and systems is crucial for ensuring efficient operation, reducing down-time, and prolonging the life of costly assets. There are state-of-the-art machine learning methods supporting this task. However, in this paper, we argue, that both efficiency and understandability can be improved by the use of explainable AI methods that analyze the importance of features used by the machine learning model. In the paper, we analyze the feature importance before a failure occurs to identify events in which an increase in importance can be observed and based on that indicate attributes with the most influence on the failure. We demonstrate how the analyses of Shap values near the occurrence of failures can help identify the specific features that led to the failure. This in turn can help in identifying the root cause of the problem and developing strategies to prevent future failures. Additionally, it can be used to identify areas where maintenance or replacement is needed to prevent failure and prolong the useful life of a system.

Keywords: explainable AI · machine learning · artificial intelligence · domain knowledge

1 Introduction

In the era of Industry 4.0, the integration of advanced technologies such as artificial intelligence (AI) and the Internet of Things (IoT) is revolutionizing the way in which industries operate. However, as these technologies become more prevalent, it is essential that they are able to provide clear and interpretable explanations for their decision-making processes. This is particularly important in the energy industry, where decisions made by AI systems can have significant consequences for the stability and sustainability of the processes.

In this paper, we aimed to demonstrate that in assets where failures are caused by component degradation the early symptoms of this degradation can be detected by the classification model which furthermore allows identifying causes of the specific failures. We used an Explainable Artificial Intelligence method (XAI), specifically the

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_33

SHAP (SHapley Additive exPlanations) algorithm [12] to identify these symptoms, indicating that the process of system degradation can be observed. To mitigate the presented challenges, we tested two approaches. First, we used a supervised learning problem to identify failures and XAI methods to verify the degradation process. Second, we used an unsupervised learning problem to identify anomalies in data and based on identifying this degradation also. To cope with that, we demonstrate our research with the use of the SHAP algorithm.

The paper is organized as follows: In Section 2 we describe the papers that cover approaches to anomaly detection and predictive maintenance. In Section 3 we present our method of detecting early symptoms of asset wear in the context of identifying areas where this wear occurs. We evaluate the method on two datasets in Section 4. Then Section 5 presents and discusses our results. Finally, in Section 6, we summarize our work.

2 Related works and motivation

Anomaly detection is a process of identifying unusual data points that do not conform to expected patterns. Two common methods for detecting anomalies are data-driven and model-based approaches [8], [13]. Data-driven methods can be further categorized into supervised and unsupervised, where supervised methods use labeled data to learn what is considered an anomaly, while unsupervised methods use techniques such as autoencoders or density-based clusterings to identify anomalies. In recent years, advances have been made in the field, such as using a CNN (convolutional neural network) to imitate human vision and decision-making for anomaly detection, or a multi-step approach that analyzes time series data in both the time and frequency domains [3], [5].

In a study [9], the authors decided to use a variation autoencoder to calculate reconstruction error, and based on the results, they labeled this error as anomalous or not. They then built a classifier that learned which points were anomalous and used it to provide explanations with the help of a SHAP algorithm [12]. However, it is worth noting that the performance of these techniques can vary depending on the specific application and the dataset available. The generalization of these methods to other contexts remains an open problem.

In [11] authors presented a new predictive maintenance policy called Sensory-updated Degradation-based Maintenance Policy (SUDM) that utilizes real-time component degradation signals and component population data to predict residual life and schedule maintenance. The policy is evaluated using a simulation model and compared with two other benchmark policies, resulting in a lower frequency of unexpected failures and lower overall maintenance costs.

Authors in [15] use autoencoder for anomaly detection instead of traditional health index to detect bearing faults. Deep neural networks extract healthy bearing representations and decoded signal residuals for fault detection. Setting an appropriate threshold for early detection is challenging without increasing false positives. Training with healthy signals is difficult to distinguish from early degradation stages.

The study [10] predicts the degradation stages of rolling-element bearings in pharmaceuticals using high-frequency vibration data. They propose a framework that uses k-means and an autoencoder to generate a labeled dataset for training a supervised model.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_33

The results are reliable and scalable, based on experiments on the FEMTO Bearing dataset.

Considering the papers presented above, many approaches for anomaly detection do not take into account the domain of the anomalies. This means that in most cases, algorithms detect many anomalies that do not actually reflect real problems with the asset. What is more, taking into account the number of detected anomalies by the machine learning algorithm we believe that the anomalies which could have an impact on the system's working conditions should characterize by the early symptoms (some degradation process). It motivated us to develop an original method that allows for explaining which anomalies are in fact crucial for the asset.

3 Feature Understanding Method

The study consists of the following steps, which are divided into two machine-learning problems. In our research, the first case involved a supervised ML problem where failure periods were available, but in the second case we focused on finding anomalies and relying on this build classifier because such periods were not available. Finally, for both cases, we used the SHAP algorithm to analyze feature importance in time.

The Shapley value is determined by evaluating the value of the feature in all possible combinations with other features and weighting and summing the results. It is defined through the value function of the features in the model as presented in Equation:$\Phi_j = \sum_{S \subseteq 1,\ldots,p/j\}}^{s} \frac{|S|!(p-|S|-1)!}{p!}(val_x(S \cup j) - val_x(S))$

It is defined by evaluating the prediction of a subset of features (S) in the model while marginalizing the features that are not included in the subset. This is done by using a value function that takes in the vector of feature values of the instance to be explained, and the number of features (p) in the model Equation where $val_x(S) = \int f(x_1, ..., x_p) d\mathbb{P}_{x \notin S} - E_X(f(X))$[14] The detailed explanation of equations is presented in [14].

Our method was evaluated on two datasets C-MAPSS [17] and real data from the steel plant which represents measurements from precess. In the first example, we used the dataset which has been already labeled [17]. So, the analyzed case was treated as a supervised problem approach. As a classifier algorithm, we used XGBoost classifier [6]. One advantage of this algorithm is its efficiency in computational time and obtained scores. The algorithm is based on decision trees, which in combination with the SHAP algorithm, is much more efficient than other classifiers that are not tree-based. In the second example, we used an unlabelled dataset where the neural network was applied to calculate the reconstruction error and obtain labels for the prediction. Then the procedure of classifier application was repeated for the labeled dataset.

3.1 Supervised task

Both approaches use a classification algorithm to train a model which later is used by the explainer algorithm. Classification is a supervised learning task, where the goal is to predict the class or category of a given data point based on a set of input features or attributes. Mathematically, it can be represented as a mapping function $f(x)$ which maps a given input x (a feature vector) to a class label y. The function $f(x)$ is learned

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_33

from a labeled training dataset $D = (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, where x_i is the i-th input feature vector and y_i is the corresponding class label [18].

In this task, two cases can be considered. Multi-class classification is a classification task with more than two possible outcomes that should be expected as an output. In the case of anomaly detection, a binary classification type of classifier is used. In this case two possible outcomes should be expected, such as predicting whether a data point is anomalous or not. In our research, we used the XGBoost classification algorithm which is based on the tree ensemble model design to be highly scalable. XGBoost builds an additive expansion (adds new models to the existing ones) of the objective function by minimizing a loss function [4], [6].

3.2 Unsupervised task

In the unsupervised task firstly we concentrate on anomaly detection. To deal with that we used an autoencoder-based model. Autoencoders are a special type of neural network that was first introduced in [16]. They are trained to recreate their input, and their main goal is to learn an informative representation of the data in an unsupervised manner. This representation can then be used for various purposes such as clustering. The issue, as defined in [1], is to learn functions. A and B presented in the following equations $A : R^n \rightarrow R^p$ and $B : B^n \rightarrow B^p$ to solve the equation $argmin_{A,B} E[\Delta(x, B \cdot A(x))]$

The expectation of the distribution of x, represented by E, is used in conjunction with the reconstruction loss function, represented by Δ. This function measures the difference between the output produced by the decoder and the original input, typically using the L2 norm [2].

Autoencoders [7] can be used as anomaly detection methods that use dimensionality reduction to try to identify a specific subspace where the normal and abnormal data differ significantly. This is done by taking a set of normal training data, represented as d dimensional vectors $x_1, x_2, ..., x_n$, $(x_i \in R^d)$ and using a model to project them into a lower-dimensional subspace. The output of this process is a set of reproduced data $x_1^{'}, x_2^{'}, ..., x_n^{'}$. The goal is to minimize the reconstruction error, which is the difference between the original and reproduced data, in order to find the optimal subspace for anomaly detection and is defined by equation $\varepsilon(x_i, x_i^{'}) = \sum_{i=1}^{d}(x_i - x_i^{'})^2$ When the data in the test dataset is similar to the typical patterns established during training, the error in reconstructing it will be lower. However, data that deviate from these patterns will have a higher reconstruction error. By setting a threshold for the reconstruction error defined by Equation, it becomes simple to identify and classify abnormal data $c(x_i) = \{^{normal \quad \varepsilon_i < \theta}_{abnormal \varepsilon_i > \theta}$ In the next section, we present an experimental evaluation of our method on two data sets.

4 Evaluation

First, we evaluated the method using a commonly used benchmark data set with synthetic data. Then, we used a real data set obtained from our industrial partners. As it is demonstrated in the remainder of this section, both experiments resulted in promising results.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_33

4.1 Experiment on the C-MAPSS dataset

The C-MAPSS dataset [17] is commonly utilized for research on predicting future performance and maintenance of systems. It includes the outcomes of a simulation of a turbofan engine utilizing the C-MAPSS software, which is provided by NASA.

C-MAPSS dataset description The dataset has information on hundreds of turbofan units with 3 operational parameters and 21 measurements taken during each unit's operation. The units deteriorated gradually over time, leading to the failure of the high-pressure compressor. The data is organized into cycles and split into four scenarios, reflecting the varying rates of deterioration and influencing factors.

To evaluate our research, we used the C-MAPSS dataset, which consists of 15631 rows and 29 columns. For our case, we decided to remove the columns like unit and cycle from the original dataset because data in this column were not directly connected with the sensor measurements. To train the classification model, we set an additional parameter in the model responsible for the weights of the classes. As a result, we obtained the F1-score 0.97 for class 0 and 0.83 for class 1 where class 0 means normal working condition and class 1 means failure.

In the Figure 1, there is presented the distribution of the SHAP values takes into account the whole features available in the dataset. For this presentation, the box plots were used where the x-axis is the cycle time. The red and green lines, there are marked respectively the maximum and minimum values of the SHAP values, which were calculated based on the aggregated data. According to previous obtain results, we can see that considering the whole features in the dataset, the SHAP values increase over time.

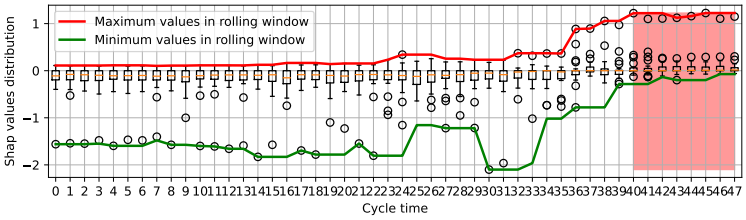

Fig. 1. SHAP values distribution during whole cycle

4.2 Experiment on hot-rolling process dataset

Our research focuses on the hot rolling process of steel at the highly automated Hot Rolling Mill (HRM) or Hot Strip Mill (HSM) at the ArcelorMittal Poland company in Krakow. The process involves heating a cold slab to around $1200°C$ in a walking beam furnace, passing it through the roughing mill (RM) to reduce thickness and width, and then through the finishing mill (FM) where the steel's thickness is further reduced. The steel is then cooled in a laminar cooler using water [9].

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_33

Hot rolling process dataset summary The dataset consists of 14,443 instances of hot rolling process data. The measurements consist of physical values of temperatures, stresses, thickness, etc. We used 35 variables to build the model, which were generated from raw values by mapping these features to a new set of values to make the data representation more relevant and easier to process for later analysis (features transformation). We used Dense Variational Autoencoder to detect anomalies. The training dataset which where provided as input to the autoencoder. The main parameters used during the training process are follwing: Latent space shape: 4, Activation function: elu, Batch size: 32, Epochs: 300, Dropout: 0.4 After learning the model, we compared the received signal from the decoder with the signal that was treated as the input to the encoder. To use the supervised classification algorithm, we had to evaluate which data points were normal and which were abnormal (anomalies). We specify a threshold that is 0.99 percentile of the reconstruction. Based on that we labeled the dataset and perform a classification algorithm with the following scores: 1.0 F1-score for class 0 and 0.72 F1-score for class 1.

Hot rolling process evaluation As a result of using the SHAP algorithm to classify the anomalies, we obtained the results, shown in Figure 2. Similarity to the examples presented in the Section 4.1 we aggregate the most important features into a single Figure, where each Figure corresponds to a different anomaly. Anomalies are marked with red rectangles or lines if an anomaly has been detected and lasted only one timestamp. In each case, we are able to observe an increase in the importance of the features before the anomaly occurred. In addition, we are able to see which features an increase in importance more and how quickly.

In the analyzed case, in three of the four graphs presented at the beginning of these anomalies, we observe an increase in the importance of temperature sensors, in two anomalies also an increase in the stress feature, and in one case an increase in torque. In order to fully evaluate these results, we asked an expert from Accerol Mittal, who was responsible for providing the data, to check the results obtained. Based on his process knowledge and experience, all temperature-related features should be reflected in the indicated anomalies because these attributes are the most important for this hot-rolling process. We choose one of the anomalies and presented it on the box plot chart in Figure 2. The maximum values and minimum values of feature importance are marked by the red and green lines respectively. However, taking into account what happens between these two lines thanks to the analysis range of box plots we can say that the range of all features' importance increases before and are respectively high during the anomaly.

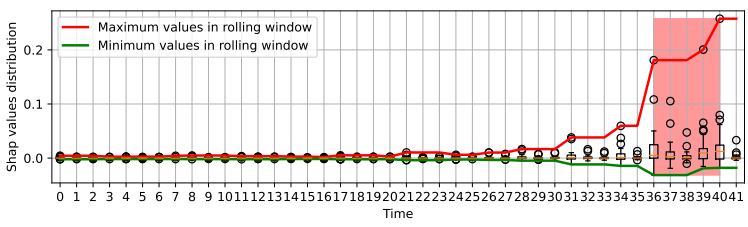

Fig. 2. SHAP values distribution before and during anomaly

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_33

5 Discussion

The work presented in this paper is based on the authors' assumption that Explainable Artificial Intelligence methods, in addition to global explanation of classification algorithms, allow to provide more information about the process and detected events in individual key events for the operation of the system.

The assumption was tested on two datasets where the events were caused by some wear processes. In both cases, we were able to identify features where the increase in the importance of features was increasing over time even before the actual event occurred. Moreover, based on this, we were able to identify which features increase the most, which allows us to better understand the behavior of the model.

The correctness of the obtained results was confirmed by an Accelor Mittal expert, who pointed out that in all the analyzed units of the hot rolling set, the temperature may be a key parameter responsible for faster wear of the rolling stage components. Given this opinion, it is reasonable to undertake further research on this topic and ultimately build an anomaly detection algorithm that takes into account the sensors which are not directly correlated with asset wear. In order to fully take into account the obtained results and expert opinions, it remains to conduct further research in order to force the model to focus on the features relevant to the wear of the element and at the same time prevent them from lowering the scores.

In addition, based on the expert's opinion and using the methods of XAI, we demonstrated that a well-prepared model built on the basis of well-prepared data is able to reproduce the actual degradation processes taking place in the plant.

6 Conclusion

In our work, we focused to analyse the causes of the detected events to validate if the anomaly characterizes early synthons. We treated these symptoms as a degradation process which led to event detection. In this work, we analyzed two examples of different cases where such behaviors can be observed. In the presented first example the events were obviously indicated by the labels. However, in the second example, these events were detected by the anomalies detection algorithm. In this example, we used an autoencoder to reproduce the original signal and find anomalies.

In both analyzed cases, we demonstrate that artificial intelligence methods are able to make predictions based on signals that are relevant to the process and can be interpreted. What's more, we were able to identify the features that are responsible for the predictions of the models, and the physical significance of these features was confirmed by an expert.

Acknowledgements This paper is funded from the XPM (Explainable Predictive Maintenance) project funded by the National Science Center, Poland under CHIST-ERA programme Grant Agreement No. $857925(NCNUMO - 2020/02/Y/ST6/00070)$

References

1. Baldi, P.: Autoencoders, unsupervised learning, and deep architectures. In: Guyon, I., Dror, G., Lemaire, V., Taylor, G., Silver, D. (eds.) Proceedings of ICML Work-

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_33

shop on Unsupervised and Transfer Learning. Proceedings of Machine Learning Research, vol. 27, pp. 37–49. PMLR, Bellevue, Washington, USA (02 Jul 2012), https://proceedings.mlr.press/v27/baldi12a.html
2. Bank, D., Koenigstein, N., Giryes, R.: Autoencoders (2020). https://doi.org/10.48550/ARXIV.2003.05991, https://arxiv.org/abs/2003.05991
3. Basora, L., Olive, X., Dubot, T.: Recent advances in anomaly detection methods applied to aviation. Aerospace **6**(11) (2019). https://doi.org/10.3390/aerospace6110117, https://www.mdpi.com/2226-4310/6/11/117
4. Bentéjac, C., Csörgo, A., Martínez-Muñoz, G.: A comparative analysis of xgboost. CoRR **abs/1911.01914** (2019)
5. Chalapathy, R., Chawla, S.: Deep learning for anomaly detection: A survey. CoRR **abs/1901.03407** (2019), http://arxiv.org/abs/1901.03407
6. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. CoRR **abs/1603.02754** (2016)
7. Chen, Z., Yeo, C., Lee, B.S., Lau, C.T.: Autoencoder-based network anomaly detection. 2018 Wireless Telecommunications Symposium (WTS) pp. 1–5 (2018)
8. Isermann, R.: Model-based fault-detection and diagnosis – status and applications. Annual Reviews in Control **29**(1), 71–85 (2005). https://doi.org/https://doi.org/10.1016/j.arcontrol.2004.12.002, https://www.sciencedirect.com/science/article/pii/S1367578805000052
9. Jakubowski, J., Stanisz, P., Bobek, S., Nalepa, G.J.: Anomaly detection in asset degradation process using variational autoencoder and explanations. Sensors **22**(1) (2022). https://doi.org/10.3390/s22010291, https://www.mdpi.com/1424-8220/22/1/291
10. Juodelyte, D., Cheplygina, V., Graversen, T., Bonnet, P.: Predicting bearings degradation stages for predictive maintenance in the pharmaceutical industry. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 3107–3115 (2022)
11. Kaiser, K., Gebraeel, N.: Predictive maintenance management using sensor-based degradation models. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on **39**, 840 – 849 (08 2009). https://doi.org/10.1109/TSMCA.2009.2016429
12. Lundberg, S.M., Lee, S.: A unified approach to interpreting model predictions. vol. abs/1705.07874 (2017), http://arxiv.org/abs/1705.07874
13. Mehdi, G., Naderi, D., Ceschini, G.F., Roshchin, M.: Model-based reasoning approach for automated failure analysis : An industrial gas turbine application (2015). https://doi.org/https://doi.org/10.36001/phmconf.2015.v7i1.2719
14. Molnar, C.: Interpretable Machine Learning (2022), https://christophm.github.io/interpretable-ml-book
15. Principi, E., Rossetti, D., Squartini, S., Piazza, F.: Unsupervised electric motor fault detection by using deep autoencoders. IEEE/CAA Journal of Automatica Sinica **6**(2), 441–451 (2019). https://doi.org/10.1109/JAS.2019.1911393
16. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning Internal Representations by Error Propagation, p. 318–362. MIT Press, Cambridge, MA, USA (1986)
17. Saxena, A., Goebel, K., Simon, D., Eklund, N.: Damage propagation modeling for aircraft engine run-to-failure simulation. In: 2008 International Conference on Prognostics and Health Management. pp. 1–9 (2008). https://doi.org/10.1109/PHM.2008.4711414
18. Sen, P.C., Hajra, M., Ghosh, M.: Supervised classification algorithms in machine learning: A survey and review. In: Mandal, J.K., Bhattacharya, D. (eds.) Emerging Technology in Modelling and Graphics. pp. 99–111. Springer Singapore, Singapore (2020)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_33