Assessment of Explainable Anomaly Detection for Monitoring of Cold Rolling Process

Jakub Jakubowski[1][0000−0002−4773−9086], Przemysław Stanisz, Szymon Bobek[2][0000−0002−6350−8405], and Grzegorz J. Nalepa[2][0000−0002−8182−4225]

[1] Department of Applied Computer Science, AGH University of Science and Technology, 30-059 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

Abstract. The detection and explanation of anomalies within the industrial context remains a difficult task, which requires the use of well-designed methods. In this study, we focus on evaluating the performance of Explainable Anomaly Detection (XAD) algorithms in the context of a complex industrial process, specifically cold rolling. We train several state-of-the-art anomaly detection algorithms on the synthetic data from the cold rolling process and optimize their hyperparameters to maximize its predictive capabilities. Then we employ various model-agnostic Explainable AI (XAI) methods to generate explanations for the abnormal observations. The explanations are evaluated using a set of XAI metrics specifically selected for the anomaly detection task in industrial setting. The results provide insights into the impact of the selection of both machine learning and XAI methods on the overall performance of the model, emphasizing the importance of interpretability in industrial applications. For the detection of anomalies in cold rolling, we found that autoencoder-based approaches outperformed other methods, with the SHAP method providing the best explanations according to the evaluation metrics used.

Keywords: machine learning · explainable artificial intelligence · predictive maintenance.

1 Introduction

In the era of Artificial Intelligence and Industry 4.0, manufacturing companies gain new opportunities for development and improvement of their processes. One of the fields that can greatly benefit from these trends is the monitoring and maintenance of the equipment in the manufacturing facilities. The digitalization of production allows companies to collect and store large amounts of data from sensors. This data can be utilized using Machine Learning (ML) methods to detect anomalies, diagnose faults, and perform root cause analysis in an online manner. All these tasks belong to a broader concept of predictive maintenance, which aims to estimate the current condition of the equipment or predict its

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_24

useful life to optimize maintenance schedules and help avoid dramatic failures, which can lead to significant losses to the company.

Our study focuses on the application of anomaly detection in the steel industry, particularly in the cold rolling process. The primary objectives of cold rolling are reducing the thickness of steel strip, improving surface finish, flatness, and increasing hardness. A typical cold rolling mill is composed of rolling stands which are placed in tandem. At each stand, the steel strip is gradually reduced to reach a target thickness at the exit of the mill. Figure 1 presents a schematic diagram of the cold-rolling process. The prediction of failures and anomalies in cold rolling processes is challanging task, due to low frequency and high diversity of the abnormalities. These issues can be addressed by anomaly detection methods, which learn the normal behavior of the process, and measure the direcptancy between the observed variables and normal working conditions. However, state-of-the-art anomaly detection methods are black-boxes, which means that predictions of these models are difficult to interpret by humans, which hinders the applicability of these methods in practice. Understanding the model's decision is vital to ensure the applicability of the method because adequate corrective actions must be taken by the crew upon detection of anomaly.

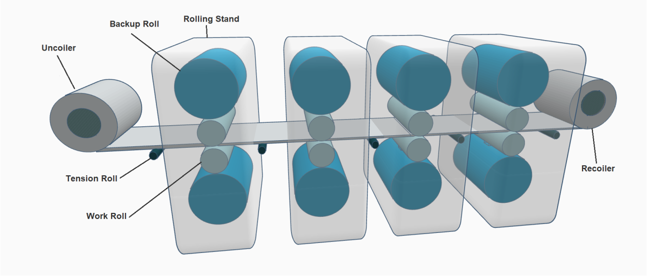

Fig. 1. Schematic diagram of 4-stand cold rolling mill

To address interpretability of black-boxes, Explainable Artificial Intelligence (XAI) has emerged, which aims to clarify the decisions of ML models for human understanding. Model-agnostic XAI methods provide explanations without prior knowledge of the model architecture, making them suitable for all types of ML models. While assessing performance of ML models is a well-established task, validating XAI methods remains a challenge. Currently, researchers often rely on human-based assessments or anecdotal examples, which do not provide a global perspective. A more robust approach involves applying metrics for quantitative comparison across different methods [24]. The selection of these metrics depends on the specific problem, allowing an assessment of the quality of the explanation.

In this paper, we assess how the selection of ML model and XAI method, can impact both the predictive performance and explanatory capabilities. To the best of our knowledge, this is the first paper that aims to quantitatively evaluate multiple XAI methods on the anomaly detection task in an industrial use-case. We propose a recipe on how to build Explainable Anomaly Detecion (XAD)

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_24

models, which account for quality of predictions and explanations. We selected several anomaly detection and XAI methods and applied them to the data from the cold rolling process. To ensure that our results are not biased by the manual labeling of the process data, we created a synthetic dataset, which simulates cold rolling of steel and generates anomalies within a fraction of observations.

The rest of the paper is organized as follows. Section 2 provides a brief overview of anomaly detection techniques, model-agnostic XAI methods, and evaluation metrics. Section 3 presents the proposed assessment methodology, including details on selected ML models and XAI methods. Section 4 contains the results of our simulations and their discussion. Section 5 concludes the work and proposes the directions for future research.

2 Related works

2.1 Anomaly Detection

Anomaly detection in industry involves identifying deviations from normal operations, which can lead to equipment failures, quality defects, or reduced performance. Early detection of anomalous behavior can bring significant benefits to manufacturing facilities. However, correctly identifying anomalies is problematic due to subjective biases. Chandola et al. [11] highlight challenges such as defining a normal region, evolving normal behavior, differences in anomaly perception across domains, availability of labeled data, and noise in the data. All of these challenges are applicable to industrial process monitoring, emphasizing the complexity of the task. The detection of anomalies in an unsupervised manner has been an extensive area of research. Wang et al. [35] groups the anomaly detection methods into density-based, statistics-based, distance-based, clustering-based, ensemble-based, and learning-based. The examples of anomaly detection methods that fall into each category are given in Table 1.

Category	**Methods**
Statisics-based	Mahalanobis Distance [23],HBOS [14]
Density-based	LOF [9]
Distance-based	kNN [19]
Clustering-based	DBSCAN[13]
Ensemble-based	IForest [21], HST [32], LODA [26]
Learning-based	AE [28], GAN [29]

Table 1. Examples of anomaly detection methods

The performance assessment of the anomaly detection methods is similar to the imbalanced classification, given the substantial difference between the number of normal and anomalous observations. An effective method is characterized by high precision and recall. Typically, there is a trade-off between precision and recall – enhancing one tends to reduce the other. Aggregating these metrics using the PRAUC or the F-score establishes a comprehensive value, considering both aspects.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_24

2.2 Explainable Artificial Intelligence

Explainable Artificial Intelligence focuses on understanding the decision-making processes of AI systems. Historically, AI research prioritized peak performance, leading to the development of complex ML models often labeled "black-boxes" [1]. These models lack interpretability, which poses challenges for human observers in tracking their decision path. The opacity of black-box models creates trust issues, as stakeholders may hesitate to rely on decisions they cannot comprehend, rendering the ML model impractical, especially for unforeseen decisions [2]. XAI not only addresses interpretability concerns, but also aids in model control and improvement by revealing decision processes and identifying potential flaws or errors in data or pre-processing [12]. In the context of anomaly detection, understanding the underlying causes of predicted anomalies is a crucial task. In complex industrial systems, relying solely on a single anomaly score while relegating its interpretation to users can be impractical.

LIME [27], SHAP [22], and Counterfactual Explanations [34] (CFE) are three prominent model-agnostic XAI methods. LIME generates explanations by approximating the behavior of a black-box model in a specific region of the input space with an interpretable model. SHAP uses a game-theoretic approach to quantify the contribution of each feature to a prediction by considering all feature combinations and their Shapeley values to determine how each feature influences the model outcome. CFE perturbs an observation with the objective of changing the model's decision, while ensuring that the generated CFE lies close to the original observation. It is defined as an optimization problem, which allows including additional constraints, e.g. the likelihood of the CFE or the number of manipulated features. The choice of the optimal XAI method may depend on the type of problem, data format, or end-user requirements.

Several studies explored the use of XAD in industrial settings. We found that many of these works rely on autoencoder architecture [15, 5, 17]. Some noteworthy works utilize Isolation Forest [4, 18], OCSVM [16, 6] and LOF [16]. In terms of explainability, we find that SHAP is widely adopted for this task [5, 15, 18, 16]. Other applied methods include rules [6, 31], CFE [17] and AcME [4]. However, of all the articles referenced, only [6] quantitatively evaluated the performance of the proposed XAI method. Moreover, none of these works compared explanations generated with different methods.

2.3 Metrics for evaluating XAI methods

Although quantifying the performance of ML models is straightforward, evaluating XAI methods remains challenging due to the subjective nature of explainability, varying stakeholder expectations, and context dependence. The absence of a clear ground truth, especially when human judgment is involved, makes defining a "correct" explanation vague. Numerous metrics attempt to quantitatively measure explanation quality, but selecting appropriate metrics is challenging, as their importance varies based on data type, use case, or stakeholders. Several studies sought to synthesize knowledge about explanation requirements [10,

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_24

24, 25]. Drawing on these works, we propose a set of metrics for evaluating the performance of XAD methods in industrial applications.

- **Faithfulness** – the primary requirement of every XAI method is to give the plausible explanation in the meaning that they are aligned with the actual model decision and expert knowledge.
- **Stability** (or robustness, continuity) – similar observations should yield similar explanations. This property ensures that a small change in the input or output of the model will not cause significant change in the explanation.
- **Compactness** (or complexity) – determines the size of the explanation measured by e.g. number of features used in the explanation. The low size of the explanation facilitates its understanding by humans.
- **Computational complexity** – it measures the time to produce explanations. It assesses the usability of the method in industrial applications rather than the quality of the explanation itself, since the explanations should be generated within a limited time.

We note that there is no coherence in terms of terminology used and similar metrics may have a different name depending on the author, e.g., stability in [10] is equivalent to continuity in [24]. We believe that this list is a good starting point to evaluate XAD methods in industrial applications.

3 Methodology

In this study, we evaluated the performance of several anomaly detection methods combined with model-agnostic XAI methods to understand how their selection affects the overall performance of the model. ML models were trained on the data from cold rolling, with the aim of predicting the anomalies within these data and generating explanations of models' decisions. The primary focus is to evaluate and compare the effectiveness of various ML and XAI methods in the context of cold rolling. We evaluate XAI methods using metrics described in Section 2.3 and identify the optimal combination of methods to build a robust and interpretable XAD model. Additionally, we analyze the influence of individual setting of XAI methods on the quality of explanations. The complete workflow is illustrated in Figure 2, with further details provided below.

3.1 Anomaly Detection and Explanation

We selected various anomaly detection algorithms for assessment. Each algorithm had its hyperparameters adjusted using Bayesian optimization [30] to maximize overall performance (measured by the PRAUC metric). Table 2 details the selected algorithms and their tuned hyperparameters.

Once the optimal architecture for each algorithm was determined, we generated explanations for all anomalous cases in the validation dataset. Explanations were not generated for normal cases as they would lack practical implications.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_24

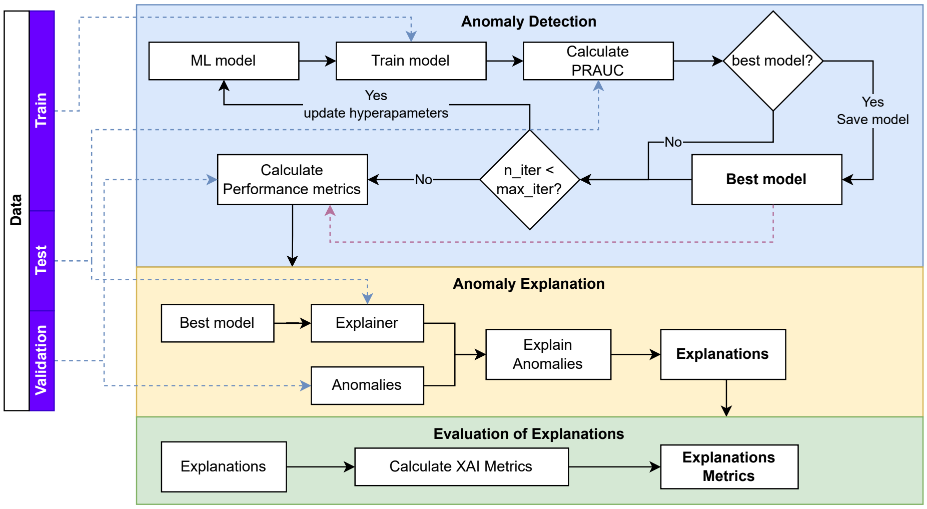

Fig. 2. Proposed methodology for evaluating Explainable Anomaly Detection models

Three model-agnostic XAI methods – SHAP, LIME, and CFE – were used for each algorithm. We applied each XAI method with different set of attributes, to observe if manipulating the method itself leads to changes in obtained results. The SHAP and LIME were used to explain the value of anomaly score, while CFE was defined as an optimization task in which we simultaniously minimize distance to the original instance, anomaly score, and sparsity penalty:

$$\min_{x}(\lambda_d \|x_0 - x\|_1 + \lambda_a f(x) + \lambda_s \|x_0 - x\|_0) \tag{1}$$

where x_0 is the explained instance, x is a counterfcatual candidate, $f(x)$ is the anomaly score of x computed by a given ML method, and λ are the weights associated with each term. To enable comparison of CFE with other methods, we convert it into feature attributions by determining the difference between the explained sample and the CFE.

3.2 Evaluation of explanations

To determine the quality of the explanations, we used various XAI metrics, which assess different aspects of the explanations. The goal of the XAI metrics is to determine quality of explanation of ML model f for observation x using the XAI method g. Firstly, we evaluate the correctness of explanations with respect to the reasoning of the model using faithfulness as proposed by Bhatt et al. [7]:

$$\mu_F(f, g; x) = \underset{S \in \binom{[d]}{[S]}}{\operatorname{corr}} \sum_{i \in S} (g(f, x)_i, f(x) - f(x_s)) \tag{2}$$

A subset d of features is randomly replaced i times with a baseline that produces perturbed samples x_s. For each sample, the difference in model output

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_24

ML method	Hyperparameters
Half-Space Trees	depth, no. trees, random state, window size
OCSVM	gamma, kernel type, nu
LODA	no. bins, no. cuts, random state
LOF	no. neighbors, algorithm type, distance metric
Isolation Forest	no. features, max. no samples, no. trees, random state
KMeans	no. clusters, random state
Autoencoder	dropout rate, hidden layers, latent layers, learning rate, no. epochs, random state
Sparse Autoencoder	dropout rate, hidden layers, latent layers, learning rate, no. epochs, beta, sparsity target, random state

Table 2. Assessed models and optimized hyperparameters

between original and perturbed observations is determined. Finally, a correlation coefficient is calculated between the attribution of the features $g(f, x)$ and these differences. Faithfulness depends on factors such as the selected baseline values, number of perturbations, number of perturbed features, and the randomness of perturbation process itself.

To evaluate the robustness of the XAI methods, we use the stability metric, based on the Lipschitz continuity, as proposed by Alvarez-Melis and Jaakkola [3]:

$$\mu_S(f, g; x) = \max_{x_j \in \mathcal{N}_\epsilon(x_i) \leq \epsilon} \frac{\|g(f, x_i) - g(f, x_j)\|_2}{\|x_i - x_j\|_2} \tag{3}$$

Stability takes samples x_j lying in the neighborhood of observation x_i, measured with Euclidean distance and constrained by ϵ. For each sample, the distance between this sample and the explained observation is determined, along with the distance between the corresponding feature attributions $g(f, x)$. Then a ratio between these values is calculated, and the maximum obtained value is the final result. Lower values of μ_S imply that explanations are more robust, as small changes in the input do not effect in drastically different explanation.

The compactness of the explanation is calculated using the entropy measure, as proposed by Bhatt et al. [7]:

$$\mu_C(f, g, x) = -\sum_{i=1}^{d} \mathbb{P}_g(i) \ln(\mathbb{P}_g(i)) \tag{4}$$

$\mathbb{P}_g(i)$ is scaled feature attribution vector $g(f, x_i)$, in a way that the sum of absolute values is equal to 1.0. Lower values of μ_C indicate a more compact explanation (using fewer features), thus enhancing its understanding by humans. Although the authors use term *complexity*, we decided to refer to it as compactness, so it is not confused with *computational complexity*.

The computational complexity of the algorithm is evaluated by measuring the time required to generate the explanations. It depends on the machine used, the ML model itself, the XAI method, and the quality of the implementation. Despite limitations, it provides valuable insight, particularly in a streaming scenario, where it is critical to ensure fast computation of explanations.

3.3 TCM dataset

All experiments are made on a synthetic dataset, which simulates the cold rolling process in a four-stand mill. The simulation is based on analytical equations describing the cold rolling process [33, 20, 8]. The data set consists of 42 variables in total, which are listed in Table 3.3.

Feature	Unit	Description
H_0, H_4	mm	thickness of coil at the entry and exit
Y_0, Y_4	mm	yield strength of coil at the entry and exit
W	mm	width of the coil
$D_1 - D_4$	mm	diameter of work rolls in each stand
$F_{t0} - F_{t4}$	kN	interstand tensions
$F_{r1} - F_{r4}$	kN	rolling force in each stand
$T_{r1} - T_{r4}$	kN	rolling torque in each stand
$V_{r1} - V_{r4}$	kN	rolling speed in each stand
$S_1 - S_4$	kN	rolling gap in each stand
$R_1 - R_4$	%	redutction in each stand
$I_1 - I_4$	kN	motor current in each stand

Table 3. Prediction perfromance of best models

Given the characteristics of the production line and its condition (which is continuously updated), the data generator randomly selects a steel coil from a pool of 20 different prodcuts. For a given product, a simulation is perfromed, which is based on several assumptions of the cold rolling process e.g. friction coefficient, tensions, reductions. At each calculation, there is a small probability that an anomalous observation will be generated. We defined four different types of anomalies: increased roll friction, reduced bearing efficiency, reduced motor efficiency, and abnormal reduction scheme, which can affect measurements from different stands. We generated 20,000 samples with an anomaly ratio of 3.0%. The data sample is presented in Figure 3. The upper plot presents the metadata of a coil, while the lower plot depicts measurements for the first stand.

4 Results

We first report the predictive performance of each model, measured on the validation set. The models were evaluated using the F1 score, G-mean, PRAUC, precision and recall and are presented in Table 4. In terms of predictive capabilities, the AE model significantly outperformed other methods. For most of the models, we observe that precision and recall are balanced, except for Half-Space Trees – in this case the recall is satisfactory, but the precision is very low, making this model useless in practice (due to high number of false alarms).

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_24

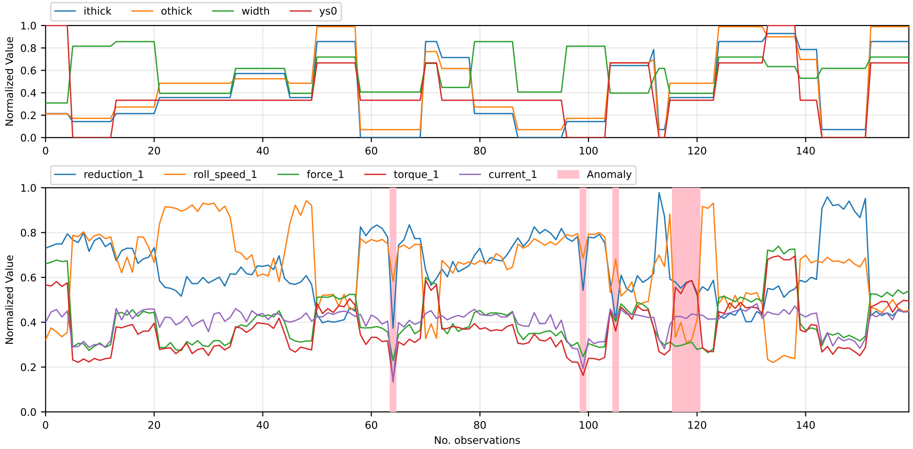

Fig. 3. Sample of TCM dataset with highlighed anomalies

Model	PRAUC	F1	G-mean	Precision	Recall
SAE	0.74	0.67	0.81	0.69	0.66
AE	**0.81**	**0.75**	**0.86**	**0.76**	**0.74**
Half Space Trees	0.06	0.08	0.54	0.04	0.40
KMeans	0.27	0.28	0.52	0.29	0.28
Isolation Forest	0.39	0.35	0.58	0.37	0.34
LODA	0.36	0.32	0.60	0.28	0.37
Local Outlier Factor	0.60	0.55	0.67	0.71	0.45
One-Class SVM	0.44	0.40	0.65	0.37	0.43

Table 4. Prediction perfromance of best models

Additionally, we evaluate the recall of the models with respect to the types of anomalies. This give us a better understanding of the capabilities of each model, the results are presented in Table 4. Each column corresponds to different type of anomaly, while the values represent the fraction of correctly detected anomalies.

Faithfulness was estimated using 100 perturbations, in which 20 features were randomly replaced with baseline values (determined based on the k-Means algorithm). This setting allowed us to obtain repeatable results. Figure 4 presents the estimation of the faithfulness metric. SHAP significantly outperformed other methods in all scenarios tested, which indicates that it is likely to be the best choice to explain anomalies in our use case. We also observe that limiting the number of base samples for SHAP from 100 to 20 did not have a negative influence on faithfulness. To validate these observations we conducted the Friedman test followed by the Nemenyi test with the p-value set to 0.05, which confirmed our hypotheses. LIME method performed decently in explaining tree-based methods, but poorly for auteoncoders. The lowest faithfullness

Model	Bearing	Electric	Reduction	Work Roll
Count	35	37	47	35
SAE	0.543	0.811	0.787	0.457
AE	**0.686**	0.892	**0.872**	**0.486**
HalfSpaceTrees	0.343	0.243	0.702	0.257
KMeans	0.086	0.081	0.702	0.143
IsolationForest	0.114	0.000	0.936	0.171
LODA	0.257	0.189	0.745	0.200
LocalOutlierFactor	0.343	0.000	**1.000**	0.314
OneClassSVM	0.143	0.297	0.915	0.257

Table 5. Recall of each model with respect to anomaly type

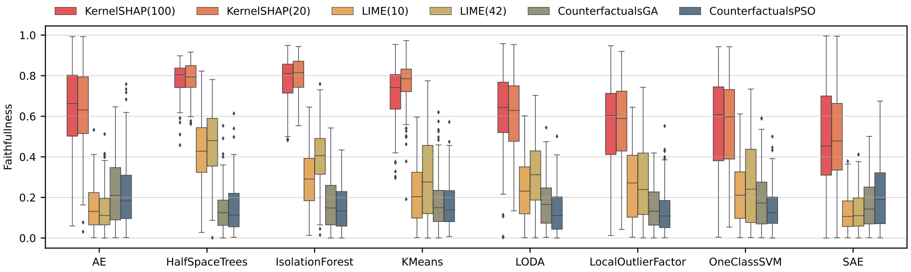

Fig. 4. Estimation of faithfulness for each combination of ML and XAI method

was achieved by CFE models, which outperfromed LIME only in models utilizing autoencoder architecture.

To determine stability, we generated 10 synthetic samples in the neighborhood of explained observations, as the anomalies lie in the low-density regions making it impossible to use samples from the data set. The comparison of stability between different XAI methods is a intricate task, due to individual characteristics of each method. To resolve the issue, we scaled all feature attributions so that the sum of their absolute values for each exlanation is equal to 1.0. Figure 5 presents the distribution of stability for all explained samples. In most cases SHAP achieved the lowest stability values, meaning that these explanations were more robust to small changes in the feature values. Again, we have performed Friedman and Nemenyi tests to confirm it, and the results indicated that statistically significant differences between SHAP and other methods were observed. Some exceptions from this behavior were observed for LIME(42), which had comparable performance to SHAP on LODA, LOF and OCSVM. Additionally, the stability of SHAP was not better than other methods for SAE. We observe that limiting the number of features in LIME greatly deteriorates the stability of this method in anomaly detection task. The obtained stability of CFE is poor, which is probably caused by the non-detererministm of the heuristics used for generating the explanations.

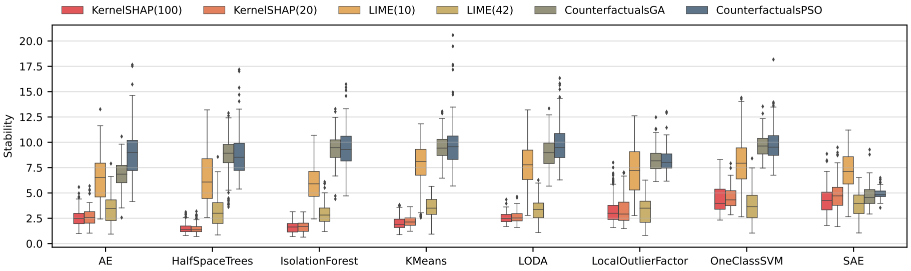

Fig. 5. Estimation of stability for each combination of ML and XAI method

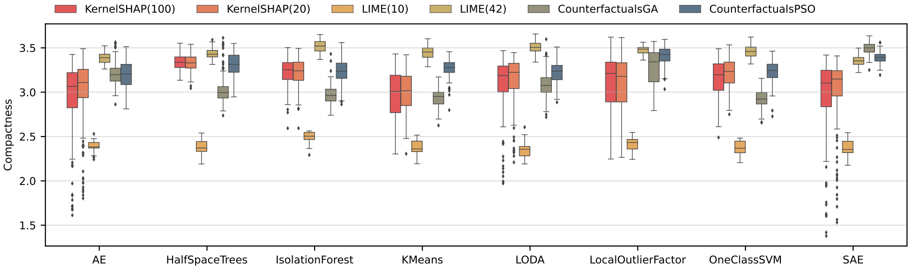

Fig. 6. Estimation of compactness for each combination of ML and XAI method

The range of values for compacntess metric spans from 0, meaning all feature attribution is assigned to a single feature, up to 3.73, which is obtained if all 42 features have equal feature attribution. Figure 6 presents the distribution of the compactness metric. We observe that most of the methods result in complex explanations, as the compactness values lie closer to the upper bound. LIME method, which was limited to 10 features, naturally obtained relatively low values compared to other approaches. Additionally, certain parts of the explanations generated for autoencoders with the SHAP method achieved satisfactory scores – these instances are related to electric motor failure. It is consistent with the reality because for this anomaly only one feature was affected.

In terms of the last of the evaluated criteria, computational complexity, we measured the time to compute each explanation. We run all experiments on the Linux machine equipped with 64-core AMD Ryzen Threadripper PRO 5995WX and 256GB of RAM. and conducted them in a sequential manner. The results are presented in Figure 7. We observe that a very high impact on the computation time has the ML model itself, which we expected, because all utilized XAI methods rely on calling the model hundreds or thousands of times to compute explanation. The shortest computation time was achieved by deep learning approaches, making them particularly useful in streaming scenarios. A significant drop in computation time is observed for the SHAP method with a reduced number of base samples. Taking into account that SHAP-based explanations

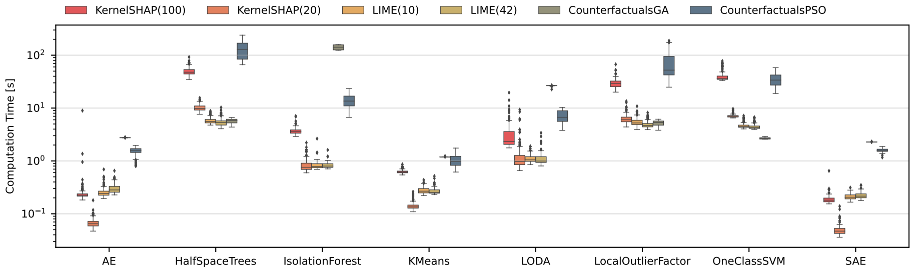

Fig. 7. Estimation of computation time for each combination of ML and XAI method

achieved comparable performance in terms of other metrics, we reason that limiting the number of base samples gives a noticeable decrease of computational complexity without a drop in performance. Regarding the CFE, we observe that this method required longest time to compute explanations, which is not reflected by the increased performance of the method. It is worth noting that there are large differences between CFE-GA and CFE-PSO, which vary depending on the model. For example, the computation time for Half-Space Trees was much shorter with CFE-GA, while in case of Isolation Forest it was the opposite. Moreover, we report that computation time varies significantly depending on the parameters chosen for the optimization methods (e.g. number of generations for GA or number of iterations for PSO). Lastly, changing the number of features in LIME does not influence the time to generate the explanations.

5 Conclusion and future works

In this paper, we evaluated different XAD methods with respect to their predictive and explanatory capabilities. We selected eight state-of-the-art anomaly detection models and three distinct XAI methods (each tested in two settings), which gave us 48 XAD models in total. All ML models were trained on a synthetic dataset, which simulates the cold rolling process of steel strip. In terms of predictive performance, the models were evaluated based on the PRAUC, F1 score, G-mean, precision and recall. To assess the performance of model explanations, we selected four distinct metrics, which considered different aspects of their explanations – faithfulness, stability, compactness, and computational complexity. The results clearly showed that the autoencoder-based models significantly outperformed other methods with respect to anomaly detection capabilities. Other methods were able to decently predict only anomalies caused by the invalid reduction scheme, which was the most complex type of anomaly (in terms of the number of perturbed features). When considering the XAI methods, we observed that SHAP significantly outperformed LIME and CFE, especially in terms of faithfulness and stability. We also note that limiting the number of base samples for SHAP did not have negative impact on its performance, but significantly reduced its computation time. This is very important in industrial applications, where data is generated at high speeds and the explanations of the anomalies should be almost instantaneous.

Despite CFE performing poorly compared to SHAP and LIME, we believe this method is worth further research, due to its high tunability. Thus, in future work, we plan to study the CFE more deeply to increase its explanatory capabilities, as we find the achieved results unsatisfying. Moreover, we plan to focus more on deep learning architectures based on the autoencoder. Additionally, we want to investigate Generative Adversial Networks, which were not considered in this study, but are known for their anomaly detection capabilities. Ultimately, we will verify our results on the data from an existing cold rolling mill, to show the practical application of the XAD.

Acknowledgements Project XPM is supported by the National Science Centre, Poland (2020/02/Y/ST6/00070), under CHIST-ERA IV programme, which has received funding from the EU Horizon 2020 Research and Innovation Programme, under Grant Agreement no 857925.

References

1. Abdullah, T.A.A., Zahid, M.S.M., Ali, W.: A review of interpretable ml in healthcare: Taxonomy, applications, challenges, and future directions. Symmetry **13**(12) (2021). https://doi.org/10.3390/sym13122439
2. Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (xai). IEEE Access **6**, 52138–52160 (2018). https://doi.org/10.1109/ACCESS.2018.2870052
3. Alvarez-Melis, D., Jaakkola, T.S.: Towards robust interpretability with self-explaining neural networks. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. p. 7786–7795. NIPS'18, Curran Associates Inc., Red Hook, NY, USA (2018)
4. Anello, E., Masiero, C., Ferro, F., Ferrari, F., Mukaj, B., Beghi, A., Susto, G.A.: Anomaly detection for the industrial internet of things: an unsupervised approach for fast root cause analysis. In: 2022 IEEE Conference on Control Technology and Applications (CCTA). pp. 1366–1371 (Aug 2022). https://doi.org/10.1109/CCTA49430.2022.9966158
5. Baek, M., Kim, S.B.: Failure detection and primary cause identification of multivariate time series data in semiconductor equipment. IEEE Access **11**, 54363–54372 (2023). https://doi.org/10.1109/ACCESS.2023.3281407
6. Barbado, A., Óscar Corcho: Interpretable machine learning models for predicting and explaining vehicle fuel consumption anomalies. Engineering Applications of Artificial Intelligence **115**, 105222 (2022). https://doi.org/10.1016/j.engappai.2022.105222
7. Bhatt, U., Weller, A., Moura, J.M.F.: Evaluating and aggregating feature-based model explanations. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. IJCAI'20 (2021)
8. Bland, D.R., Ford, H.: The calculation of roll force and torque in cold strip rolling with tensions. Proceedings of the Institution of Mechanical Engineers **159**(1) (Jun 1948)

DOI: 10.1007/978-3-031-63775-9_24

9. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: Lof: Identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. p. 93–104. SIGMOD '00, Association for Computing Machinery, New York, NY, USA (2000). https://doi.org/10.1145/342009.335388
10. Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: A survey on methods and metrics. Electronics **8**(8) (2019). https://doi.org/10.3390/electronics8080832
11. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Comput. Surv. **41**(3) (jul 2009). https://doi.org/10.1145/1541880.1541882
12. Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., Ranjan, R.: Explainable ai (xai): Core ideas, techniques, and solutions. ACM Comput. Surv. **55**(9) (jan 2023). https://doi.org/10.1145/3561048
13. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. p. 226–231. KDD'96, AAAI Press (1996)
14. Goldstein, M., Dengel, A.: Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. KI-2012: poster and demo track **1**, 59–63 (2012)
15. Ha, D.T., Hoang, N.X., Hoang, N.V., Du, N.H., Huong, T.T., Tran, K.P.: Explainable anomaly detection for industrial control system cybersecurity. IFAC-PapersOnLine **55**(10), 1183–1188 (2022). https://doi.org/10.1016/j.ifacol.2022.09.550, 10th IFAC Conference on Manufacturing Modelling, Management and Control MIM 2022
16. Hermansa, M., Kozielski, M., Michalak, M., Szczyrba, K., Wrobel, L., Sikora, M.: Sensor-based predictive maintenance with reduction of false alarms; a case study in heavy industry. Sensors **22**(1) (2022). https://doi.org/10.3390/s22010226
17. Jakubowski, J., Stanisz, P., Bobek, S., Nalepa, G.J.: Roll wear prediction in strip cold rolling with physics-informed autoencoder and counterfactual explanations. In: 2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA). pp. 1–10. IEEE (10 2022). https://doi.org/10.1109/DSAA54385.2022.10032357
18. Kim, D., Antariksa, G., Handayani, M.P., Lee, S., Lee, J.: Explainable anomaly detection framework for maritime main engine sensor data. Sensors **21**(15) (2021). https://doi.org/10.3390/s21155200
19. Knorr, E.M., Ng, R.T.: Algorithms for mining distance-based outliers in large datasets. In: Proceedings of the 24rd International Conference on Very Large Data Bases. p. 392–403. VLDB '98, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1998)
20. Lenard, J.G.: 9 - tribology. In: Lenard, J.G. (ed.) Primer on Flat Rolling (Second Edition), pp. 193–266. Elsevier, Oxford, second edition edn. (2014). https://doi.org/10.1016/B978-0-08-099418-5.00009-3
21. Liu, F.T., Ting, K.M., Zhou, Z.: Isolation forest. In: 2008 Eighth IEEE International Conference on Data Mining. pp. 413–422 (2008). https://doi.org/10.1109/ICDM.2008.17
22. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc. (2017)
23. Mahalanobis, P.: On the generalised distance in statistics. Proceedingsof the National Institute of Sciences of India **2**, 49–55 (1936)

24. Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., Seifert, C.: From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. ACM Comput. Surv. **55**(13s) (jul 2023). https://doi.org/10.1145/3583558
25. Oblizanov, A., Shevskaya, N., Kazak, A., Rudenko, M., Dorofeeva, A.: Evaluation metrics research for explainable artificial intelligence global methods using synthetic data. Applied System Innovation **6**(1) (2023). https://doi.org/10.3390/asi6010026
26. Pevný, T.: Loda: Lightweight on-line detector of anomalies. Machine Learning **102**(2), 275–304 (jul 2015). https://doi.org/10.1007/s10994-015-5521-0
27. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 1135–1144. KDD '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939778
28. Sakurada, M., Yairi, T.: Anomaly detection using autoencoders with nonlinear dimensionality reduction. In: Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis. p. 4–11. MLSDA'14, Association for Computing Machinery, New York, NY, USA (2014). https://doi.org/10.1145/2689746.2689747
29. Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: Niethammer, M., Styner, M., Aylward, S., Zhu, H., Oguz, I., Yap, P.T., Shen, D. (eds.) Information Processing in Medical Imaging. pp. 146–157. Springer International Publishing, Cham (2017)
30. Snoek, J., Larochelle, H., Adams, R.P.: Practical bayesian optimization of machine learning algorithms. In: Pereira, F., Burges, C., Bottou, L., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 25. Curran Associates, Inc. (2012)
31. Steenwinckel, B., De Paepe, D., Vanden Hautte, S., Heyvaert, P., Bentefrit, M., Moens, P., Dimou, A., Van Den Bossche, B., De Turck, F., Van Hoecke, S., Ongenae, F.: Flags: A methodology for adaptive anomaly detection and root cause analysis on sensor data streams by fusing expert knowledge with machine learning. Future Generation Computer Systems **116**, 30–48 (2021). https://doi.org/10.1016/j.future.2020.10.015
32. Tan, S.C., Ting, K.M., Liu, T.F.: Fast anomaly detection for streaming data. In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two. p. 1511–1516. IJCAI'11, AAAI Press (2011)
33. Venkata Reddy, N., Suryanarayana, G.: A set-up model for tandem cold rolling mills. Journal of Materials Processing Technology **116**(2–3), 269–277 (Oct 2001). https://doi.org/10.1016/s0924-0136(01)01007-x
34. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv. JL & Tech. **31**, 841 (2017)
35. Wang, H., Bah, M.J., Hammad, M.: Progress in outlier detection techniques: A survey. IEEE Access **7**, 107964–108000 (2019). https://doi.org/10.1109/access.2019.2932769