Baseline Selection for Integrated Gradients in Predictive Maintenance of Volvo Trucks' Turbocharger

Nellie Karlsson
School of Information Technology
Halmstad University
Halmstad, Sweden
email: nelkar17@student.hh.se

My Bengtsson
School of Information Technology
Halmstad University
Halmstad, Sweden
email: mypet17@student.hh.se

Mahmoud Rahat
Center for Applied Intelligent Systems Research (CAISR)
Halmstad University
Halmstad, Sweden
email: mahmoud.rahat@hh.se

Peyman Sheikholharam Mashhadi
Center for Applied Intelligent Systems Research (CAISR)
Halmstad University
Halmstad, Sweden
email: peyman.mashhadi@hh.se

***Abstract*—The new advances in Vehicular Systems and Technologies have resulted in a sheer increase in the number of connected vehicles. These connected vehicles use IoT technologies to communicate operational signals with the OEMs, such as the vehicle's speed, torque, temperature, load, RPM, etc. These signals have provided an unprecedented opportunity to adaptively monitor the status of each piece of the vehicle's equipment and discover any possible risk of failure before it happens. This emerging field of study is called predictive maintenance (also known as condition-based maintenance) and has recently received much attention. In this paper, we apply Integrated Gradients (IG), an XAI method until now primarily used on image data, on datasets containing tabular and time-series data in the domain of predictive maintenance of trucks' turbochargers. We evaluate how the results of IG differ, in these new settings, for various types of models. In particular, we investigate how the change of baseline can affect the outcome. Experimental results verify that IG can be applied successfully to both sequenced and non-sequenced data. Contrary to the opinion common in the literature, the gradient baseline does not affect the results of IG significantly, especially on models such as RNN, Long Short Term Memory (LSTM), and GRU, where the data contains time series; the effect is more visible for models like MLP with non-sequenced data. To confirm these findings, and to understand them deeper, we have also applied IG to SVM models, which gave the results that the choice of gradient baseline has a significant impact on the performance of SVM.**

***Index Terms*—Explainable AI (XAI), Predictive Maintenance, Integrated Gradients, Machine Learning.**

I. Introduction

With the increase in popularity of artificial intelligence, several challenges have been brought to light, for example, the lack of transparency, debugging difficulty, lack of control, and biased outcomes that may not represent the real world with its principles and norms [1]. Even though AI is a powerful tool for predictions, it does lack transparency. A significant reason for this is the black-box structure that comes with deep learning methods such as Deep Neural Networks (DNNs), where their hidden layers are hard to visualize for human understanding. In contrast to DNNs and other similarly complex AI models, there are several simpler approaches that are more interpretable and easier to visualize, for example, decision trees; however, they often have limited accuracy [2]. The trade-off between explainability and accuracy can therefore be a challenge.

Because of AI's lack of transparency, it can be challenging to trust the important life-changing decisions the algorithms may take. The AI algorithms and methods give us an answer, but not a *why* or a *how* to that answer. It is hard to trust an algorithm without knowing why and how it made a specific decision. As a result of these challenges, the subject of Explainable Artificial Intelligence (XAI) has arisen. Even though the interest in XAI has increased in the last few years, the term XAI was first coined by Van Lent et al. in 2004 [3]. However, the concept of explainability in machine learning has existed since the 1970s according to [2]. Today, one of the goals of XAI is for humans to understand and trust the reason behind the decisions of an AI model while the model maintains a high prediction accuracy. The theory behind XAI can usually be simplified and divided into four main principles: to justify, to control, to improve, and to discover [2]. This is also a goal for this paper: to explain the reasoning behind the resulting predictions in an understandable way.

There are already several techniques to use for XAI of different kinds, for example, scope-related and/or model-related. Scope-related techniques are divided into two categories: global and local interpretability. Global interpretability is when the technique follows the whole reasoning leading to all of the predictions of the chosen model and understanding its logic. However, global interpretability can be hard to achieve

 ISBN: ISBNFILL

in practice, mainly when it comes to machine learning (ML) models with a large number of parameters. Local interpretability is easier to implement in reality, considering its main focus is on explaining a single prediction and not several. Local interpretability is also the primary approach for the explainability of predictions made by deep neural networks (DNNs) [2].

This paper focuses on explainable AI for Predictive Maintenance (PdM). PdM is a condition-driven preventive method and is used to improve the productivity of a machine by regularly monitoring the parts of the machine to avoid a run-to-failure approach or to maintain a healthy machine [4]–[6].

We have performed several experiments on a real-world turbocharger system dataset provided by Volvo, which is used to predict the remaining useful life. The data is a time-series dataset containing over 400 sensor values and on average 20 timestamps sampled biweekly [7]. To be able to more accurately evaluate the impact of the gradient baseline of integrated gradients (IG) in predictive maintenance, we resort to simulated data. This is due to the fact for the chosen simulated data the feature importance is known to the research community and accordingly easier to evaluate and justify. The first simulated dataset is the Turbofan Engine Degradation Simulation Data Set (CMAPS), which is run-to-failure data that could be used to predict remaining useful life. The dataset contains time as well as sensors reading, which makes the dataset similar to the Volvo dataset [8]. The Tennessee Eastman Process Simulation dataset (ETEPS-CP) is used for the second simulated dataset. The dataset contains information about chemical plants, where some features are measurements while some are manipulated values. This dataset is used as a comparison since it is known which features are measurements and should have more impact on the predictions. The target for the dataset is set to be a classification problem since it is known that the chemical plant runs normally until a fault is induced[[9], [10]]. The dataset contains 54 sensor values and when transformed into time series it contains approximately 38000 timestamps.

A significant difficulty in implementing Integrated Gradients is determining the gradient baseline, which plays an important part in the results. When using images as inputs, it is most common to use a black or white image as a gradient baseline, but the choice of the gradient baseline is not as clear for tabular data. The gradient baseline for tabular data varies depending on the dataset type, and there is not much research on finding the optimal gradient baseline for these types of datasets. We, therefore, want to find a systematic way of defining the optimal gradient baseline for integrated gradients with tabular data as input. The meaning of the word *baseline* in this report refers to the baseline used in integrated gradients, which is explained in more detail in Section III.

This paper explores the following:

1) How the baseline for integrated gradients can be chosen for tabular data in predictive maintenance.
2) How the gradient baseline affects the outcome of different models.

II. Related Work

There are different types of machine learning algorithms and explainable AI that have been used for predictive maintenance. Some algorithms and work that have been adapted to predictive maintenance are *Bagged trees ensemble*.

In the work done by [11], they have used bagged tree ensemble, decision trees and normalized feature deviations to get the model more interpretative. The result of their work concluded that when using bagged trees ensemble, the decision trees as an explanation got a higher quality but did not generate a complete explanation on all test cases as proposed to normalized feature deviations, which got a lower quality of explanation but generated consistent explanations.

In addition to the work described above, the paper [12] evaluates the previous work as well as added LIME to interpret the result. In the paper, they used *Random Under Sampling* (RUS) and boosted trees ensemble, which successfully and correctly classifies all failures as a comparison where the bagged trees ensemble did not.

Other machine learning algorithms used for predictive maintenance and anomaly detection are *Principal Component Analysis* (PCA), *null-space*, *One-Class Support Vector Machines* (OC-SVM), *Extreme Learning Machine* (ELM), and *2 Dimensional Convolutional-based Neural Network Autoencoder* (2D-CNN-AE), which are compared in [13]. The paper compares the different approaches by using the F1-score, where the best approaches are concluded to be the null space and 2D-CNN-AE. Due to the capability of 2D-CNN-AE to detect even small failures, it is outperforming the other methods.

To handle time-series data in prediction, a combination of Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) is proposed in paper [14]. The paper predicts the remaining lifetime of aircraft engines by comparing Convolutional Neural Network (CNN), LSTM, and a combination of the two algorithms, CLSTM, where the result is that the combination of the two algorithms provides the highest accuracy rate. Another work that handles both CNN and LSTM together is a fault diagnostic system on mechanical data from a gearbox [15]. The result of the algorithm proposed is 97% accuracy, where the algorithm is able to detect which fault is detected.

III. Integrated Gradients

Integrated Gradients (IG) is a technique for model interpretability used to visualize the relationship between the prediction of the model and the input features, often when the input is an image. Similar to SHAP, IG is also inspired by game theory, especially the Aumann-Shapley value, which SHAP is based on [16]. Below, the way to compute IG is shown as well as in eq. 1:

1) The first step is to identify the input and output. In this study, the input is the sequential data whereas the last layer of the model is the output.
2) To be able to identify features that are important to the prediction of the neural network, the second step is to choose a gradient baseline as an input.

 ISBN: ISBNFILL

3) The third step is then to interpolate the chosen gradient baseline for a number of steps. The number of steps is a hyperparameter and represents the number of steps needed for the given input in the gradient approximation, where the recommended number is between 20-1000 steps,
4) After the gradient baseline has been interpolated they are preprocessed, and then a forward pass is done.
5) Lastly, the gradients for the interpolated data points are obtained and then by using the trapezoidal rule, the gradients integral is approximated.

$$IG_i^{approx}(x) :== (x_i - x_i^{'}) \times \sum_{k=1}^{m} \frac{\partial F(x^{'} + \frac{k}{m} \times (x - x^{'}))}{\partial x_i} \times \frac{1}{m} \quad (1)$$

Where:

$[x_i]$ = Input Data
$[x_i^{'}]$ = Gradient Baseline
$[m]$ = Number of Steps in the Integral Approximation

There are several advantages of integral gradients. An example is sensitivity, which means that it will give a non-zero attribution every time there is a difference in one feature between input and gradient baseline but also a difference in predictions. Another example is the invariance of implementation, where two models' feature attributions will be the same if they both are functionally equivalent, without regard to the network architecture [17].

IV. Methodology

A. Setup

Before the experiments are performed, the data needs to be prepared, which is done differently among the data sets. For the CMAPS dataset, standard scaling and hyperparameter tuning are used. For the ETEPS-CP dataset, standard scaling was used to preprocess the data. For the Volvo dataset, more preparations need to be taken, where the first step is to handle the NAN values by imputation with the mean values of each column. It is important to note that Mean Imputation (MI) can lead to biased estimates and predictions, especially if the number of NAN values is significant. Before the imputation, the dataset only contained approximately 2% of NAN values, and therefore we decided that MI is a suitable type of imputation in this dataset.

After MI, the columns containing objects or strings are label-encoded so that the model used only has float or integer as input values. Most of the ML models do not take strings as inputs, which is why label encoding is needed.

Lastly, the data needs to be scaled and normalized when employing deep learning models. For this, MinMaxScaler and standardScaler from the scikit-learn library are used.

B. Machine Learning Methods

To be able to evaluate the XAI methods, as well as the gradient baselines, we first need to implement reliable models. The models used in the experiments are Recurrent Neural Network (RNN), Long short-term memory (LSTM), Gated Recurrent Unit (GRU), and Multilayer Perceptrons (MLP). Support Vector-Machine (SVM) is also implemented to affirm the results we receive from the experiments on MLP, which can be seen in section V. These regressor models are evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). For the classification models, the measurements will be on accuracy as well as Area Under the Receiver Operating Characteristic Curve (ROC AUC) where it should present better than randomly choosing a class, which is better than 0.5. The models perform better than the baseline, which means that for the regression part, the mean absolute error is smaller than the mean value for the test sets. For the classification, the accuracy is above 50%, and the ROC AUC is over 0.5.

C. Systematic Choice of Baseline

The choice of a gradient baseline has a large impact on the result of Integrated Gradients, and it is therefore crucial to have an appropriate gradient baseline. In figure 1 we provide a systematic approach to the choice of gradient baseline, which is explained in more detail in the following paragraphs.

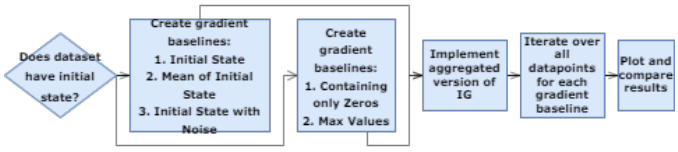

Fig. 1: A systematic approach on how to select gradient baseline.

For the Volvo dataset and the simulated datasets, the gradient baselines used are a gradient baseline with the Initial State (IS), the Mean Initial State (MIS), and the Initial State with Gaussian Noise (ISN). For datasets without an initial state, gradient baselines consisting of either only zeros or max values can be used to resemble an image. Our approach of using a gradient baseline with only zeros simulates a white image, and a gradient baseline with the max value for all features simulates a black image.

An aggregated version of integrated gradients has been used to get an overall view of the attributions. The aggregated model of integrated gradients is an iterative type of integrated gradients, where all attributions for all data points are iterated. The reason for using an aggregated integrated gradient is to get an average of all feature importance.

For all datasets, the aggregated version of integrated gradients was iterated over all data points for the different gradient baselines. The different baselines used are the initial state of the data, the initial state with gaussian noise as well as the mean value of all the states. The initial state is used as a ground truth of the datasets, where the different variations of the ground truth are used to explore the result when using different gradient baselines.

Lastly, the results between the different gradient baselines are plotted and compared, where the gradient baseline with

 ISBN: ISBNFILL

the most logical results (in accordance with domain experts) is the most favorable choice of baseline.

V. RESULTS

A. Gradient Baseline

The main experiment is to find the optimal gradient baselines for sequential and tabular datasets. We use the two simulated benchmark datasets (ETEPS-CP and CMAPS) to test gradient baselines similar to the ones used on the Volvo dataset.

1) Simulated Data - CMAPS: Figure 2 shows the features that appeared as the top three features with the most importance according to IG for all of the RNN models. In Figure 2, we can see that sensor 11 plays a significant role in the prediction of all models, as it almost always is the third most important feature. We can also see that sensor 4 is clearly an important feature for all of the models, as it rather often placed as the second most important feature according to IG.

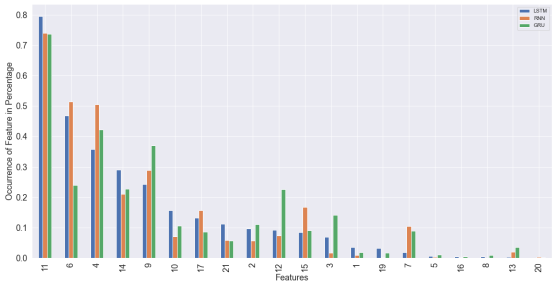

Fig. 2: Feature occurrence for top three features with the most importance in percentage according to the results of IG. These results are for the models LSTM, RNN, GRU on the CMAPS dataset, and depend on the model.

Overall, all RNN models (RNN, LSTM, and GRU) seem to give similar results when using integrated gradients, where sensors 11, 4, 6, and 9 have the most significant impact on the prediction. Looking at Figure 3 with the MLP model, we also receive sensors 11 and 4 at the top. Sensors 6 and 9, however, can only be seen towards the far right of the graph, occurring under 5% as the top 3 features.

Looking at Figure 3, the results between the gradient baselines of the MLP model also vary more than between the RNN models. Here, we can see that the gradient baselines pay a different amount of attention to a larger variety of sensors than the RNN models do in Figure 2. We can also see this in Figures 4, 5 and 6, where how large of an impact that the different gradient baselines have on each model.

In the case of the CMAP dataset, it could be that the MLP model is sensitive toward specific sensors, such as sensors 6 and 9 (as explained above). This sensitivity could lead to the baselines playing a much more significant role in the results of IG.

Another reason the results between the baselines differ much more for the MLP model than for the RNN models is that the baseline may not have as much significance when using models specified toward sequenced data. We, therefore, theorize that when using a model such as RNN where the data is sequenced, the baseline does not affect the outcome of IG to a large extent, as long as the baseline is a reasonable one (for example, the initial state). For models where the input is not sequenced, the gradient baseline affects the results of IG more.

To endorse this theory, we applied IG to a Support Vector Machine (SVM) model where the input is not sequenced. In Figure 7, we can see that the baselines play a significant role in the outcome of IG, similar to the results of MLP.

2) Simulated Data - ETEPS-CP: The results for the MLP model can be seen in Figure 10. The initial state is almost identical to the initial state with noise. However, the mean of the initial states gives completely different results. The similarities between the initial state and the initial state with noise could be that adding noise to the baseline does not change the baseline significantly or that these features are highly correlated to the prediction. When comparing to Figure 10, it is seen that the initial state and initial state with noise is similar, which also was seen in the table. Moreover, the mean initial state pays attention to a broader number of features, with a smaller number of features occurring more than others. The conclusion to draw from the dataset with an MLP model is that the initial state and initial state with noise perform better since fewer features occur in the top for around 50% of all data points.

In Figure 10, it is shown that the initial state and initial state with noise pay attention to the same features while the mean initial state pays attention to multiple features.As seen in Figure 10, the gradient baseline MIS is not shown in the figure. This is because the values from MIS are giving feature values near zero for the MLP model. When looking at the other gradient baselines, it is clear that there is no difference between the values for the feature importance.

The results for the GRU model can be seen in Figure 13. All three baselines are similar to each other in both placement and occurrence.In Figure 13, the values for the top features are similar for every gradient baseline. Compared to the MLP-model, Figure 10, which has more features occurring at the top, the GRU model gives fewer features with no difference between baselines.

The results for the LSTM model can be seen in Figure 12. However, the different features do not appear in the same placement or occurrence. In Figure 12, the values for the top features is similar for every gradient baseline as for the GRU model, Figure 13.

The results for the RNN model can be seen in Figure 11. As seen in Figure 11, the different baselines are almost identical to each other. Further looking into both GRU in Figure 13 and LSTM in Figure 12, the same pattern reoccurs, where all the different baselines give almost identical results for the same model and baselines.

However, comparing the different time-series models seen in Figure 8, the models pay attention to the same features but

 ISBN: ISBNFILL

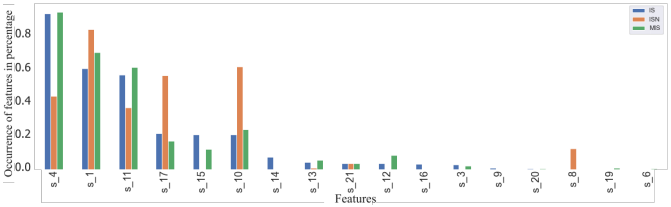

Fig. 3: Feature occurrences for top three features with the highest importance (in percentage), according to the results of IG, for the four different types of networks. These results are for the CMAP dataset for MLP model, and depend on the baselines.

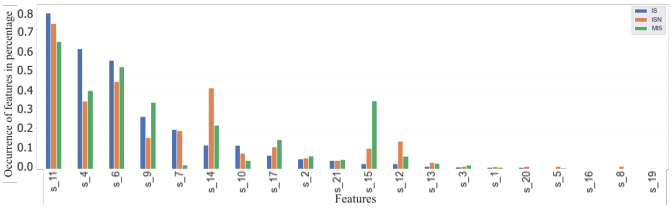

Fig. 4: Feature occurrences for top three features with the highest importance (in percentage), according to the results of IG, for the four different types of networks. These results are for the CMAP dataset for RNN model, and depend on the baselines.

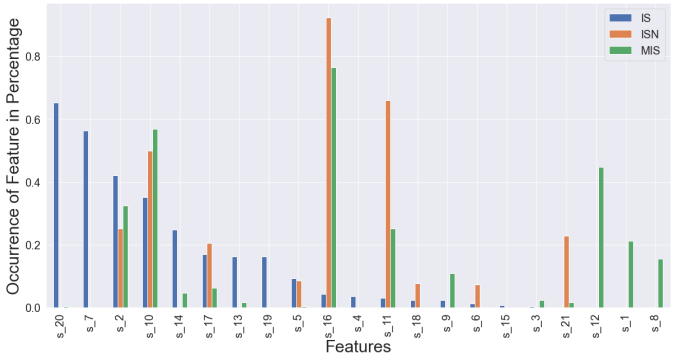

Fig. 7: Feature occurrence for top three features with most importance in percentage according to the results of IG. These results are for the model SVM on the CMAP dataset, and depend on the models.

with different magnitudes.

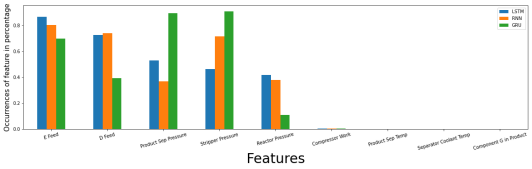

Fig. 8: Feature occurrence for top three features with most importance in percentage for IG. These results are for the models RNN, GRU and LSTM on the ETEPS-CP dataset.

The result for IG on the sequential data is once again tested as in V-A1, where similar results are presented regarding sequential data. To see if the theory that the gradient baseline affects the results of IG on non-sequential models more than sequential models also applies to the ETEPS-CP dataset, we apply SVM classification. The result of the SVM Classification

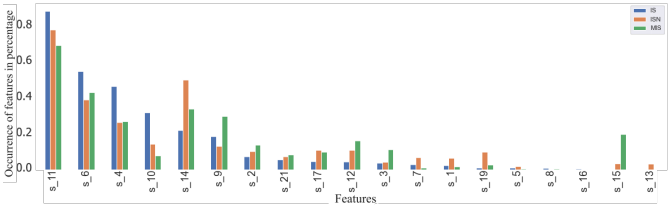

Fig. 5: Feature occurrences for top three features with the highest importance (in percentage), according to the results of IG, for the four different types of networks. These results are for the CMAP dataset for LSTM model, and depend on the baselines.

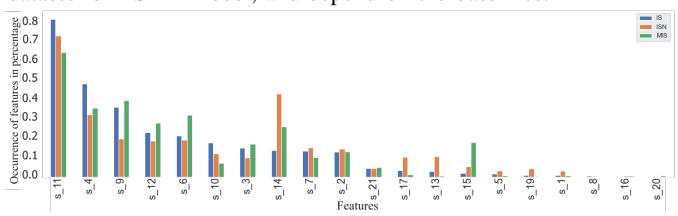

Fig. 6: Feature occurrences for top three features with the highest importance (in percentage), according to the results of IG, for the four different types of networks. These results are for the CMAP dataset for GRU model, and depend on the baselines.

can be seen in Figure 12 which strengthens the belief that the non-sequential model is more affected by the gradient baseline.

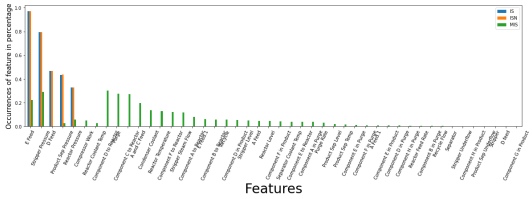

Fig. 9: Feature occurrence for top three features with the most importance in percentage according to the results of IG. These results are for the SVM Classification model on the ETEPS-CP dataset, with three different baselines.

In Figure 8, the feature occurrence from IG for LSTM, GRU, and RNN models can be seen. From Figure 8, it is shown how the different models give similar results to the IG.

The overall conclusion for the ETEPS-CP dataset is that the choice of baselines does not seem to affect the explanation for the time-series models. Moreover, when comparing the MLP to time-series, the gradient baseline significantly impacts the explanations.

3) Volvo Dataset: The results for the MLP model can be seen in Figure 14.

In Figure 14 the occurrence of the features can be seen. As with the simulated datasets, MLP pays attention to multiple features depending on the gradient baseline. Looking into the values in Figure 14, it can be seen that the initial state and initial state with noise are similar. However, the mean initial state pays attention to a broader number of features.

Evaluating Figure 14, shows that they both often occur and have a large impact on the prediction. For the MLP model,

 ISBN: ISBNFILL

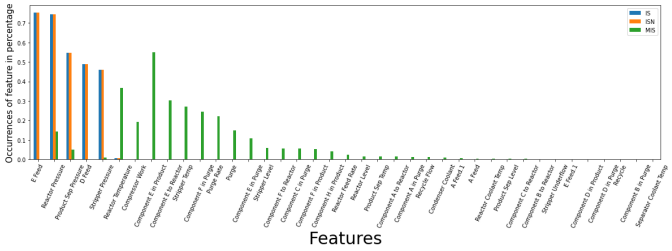

Fig. 10: Feature occurrence for top three features with the most importance in percentage according to the results of IG. These results are for the ETEPS-CP dataset for MLP model, and depend on the baselines.

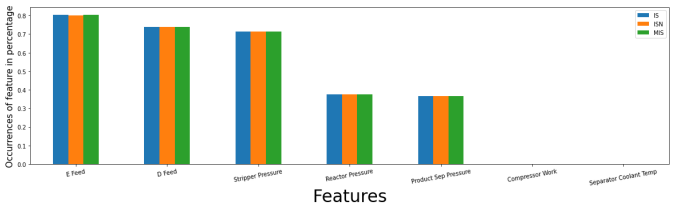

Fig. 11: Feature occurrence for top three features with the most importance in percentage according to the results of IG. These results are for the ETEPS-CP dataset for RNN model, and depend on the baselines.

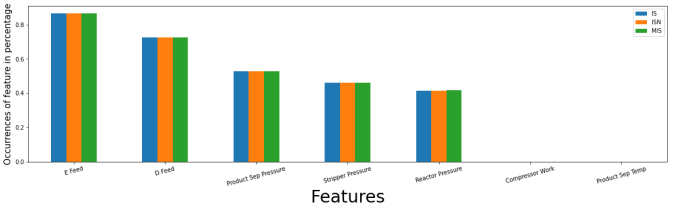

Fig. 12: Feature occurrence for top three features with the most importance in percentage according to the results of IG. These results are for the ETEPS-CP dataset for LSTM model, and depend on the baselines.

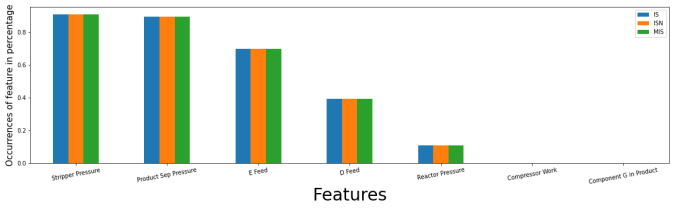

Fig. 13: Feature occurrence for top three features with the most importance in percentage according to the results of IG. These results are for the ETEPS-CP dataset for GRU model, and depend on the baselines.

the initial state is similar to the initial state with noise, while the mean initial state has a broader view of features.

The MLP-model, as seen in Figure 14, shows a broad spectrum of features where only the mean initial state has features that occur in the top three at more than 50% of the dataset. Some similarities can be seen between the initial state and the initial state with noise. However, the other datasets are not identical, which could result from having a more complex dataset. IG is also very sensitive to noise, leading to the difference between the IG for different datasets.

When comparing the baselines for the RNN model, as seen in Figure 15, the different baselines do not differ as much as for the MLP. However, the different baselines do not seem to have a huge impact on important features. By looking into GRU, Figure 17 as well as LSTM, Figure 16 the same patterns are occurring. For all time-series data there are not any significant features that occur in the top three.

The overall conclusions that can be drawn from this are how sensitive IG is towards the noise and that the baselines do not significantly impact time-series data, which is shown in sections V-A1 and V-A2. Some features occur in multiple gradient baselines; however, no feature occurs in all three different baselines for the MLP-model.

The results for the RNN model can be seen in Figure 15. IG gradient baseline initial state and initial state with noise are similar. In contrast, the mean initial state pays attention to more features, which means that the theory of gradient baseline can be applied when the dataset is robust and does not have noise since IG is sensitive to noise.

The results for the GRU model can be seen in Figure 17.

Evaluating Figure 17, it can be seen that for all gradient baselines the model pays attention to multiple features depending on the baseline. Since the data contains noise, the result of the variance of the gradient baselines can be disregarded for the outliers. However, when looking closer at the figures, some of the features occur in all baselines which strengthen the belief that the gradient baseline for sequenced data with a small amount of noise result in similar conclusions.

The results for the LSTM model can be seen in Figure 16.

When looking at Figure 16, it can be seen that the data contains noise due to some features appearing in the top ten for only one gradient baseline.

VI. Conclusions & Future Work

In this paper, we have dived deeper into the XAI method integrated gradients to justify the predictive maintenance results for a dataset provided by Volvo. To our knowledge, IG is a method commonly used for data of images and not the time-series data that the Volvo dataset contains. The lack of work done on time-series data for IG can be because the choice of gradient baseline can be seen as complex. Therefore, we have focused the paper on how to find a good baseline and how the baseline affects the result depending on the deep learning model and the type of data. We have also investigated other types of XAI methods to either justify or compare the results of our experiments.

We observed that integrated gradients are a good method to interpret the behavior of deep learning models in predictive maintenance, especially for time series data. However, we still believe that the choice of gradient baseline continuous to be seen as difficult. As stated before, we theorize that the gradient baseline's effect on the results of integrated gradients decreases for RNN models with sequential data and increases on models like MLP with non-sequential data. However,

 ISBN: ISBNFILL

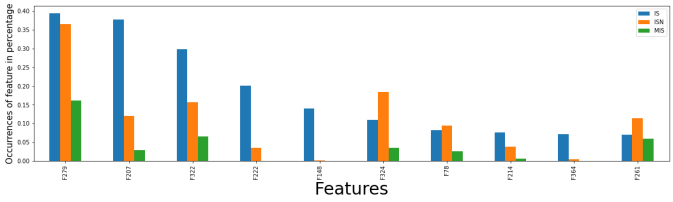

Fig. 14: Feature occurrence for top three features with the most importance in percentage according to the results of IG. These results are for the Volvo dataset for MLP model, and depend on the baselines.

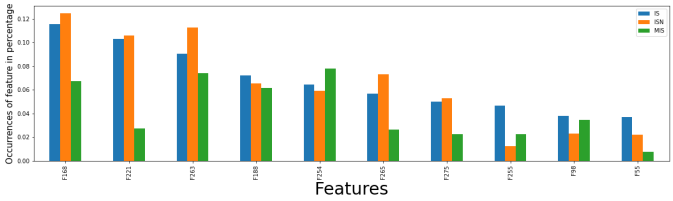

Fig. 15: Feature occurrence for top three features with the most importance in percentage according to the results of IG. These results are for the Volvo dataset for RNN model, and depend on the baselines.

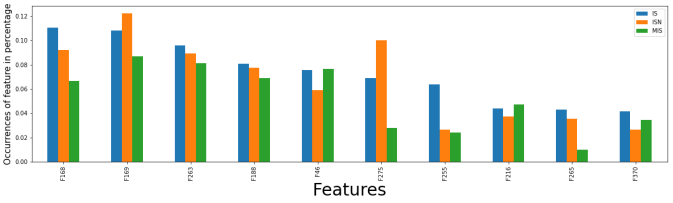

Fig. 16: Feature occurrence for top three features with the most importance in percentage according to the results of IG. These results are for the Volvo dataset for LSTM model, and depend on the baselines.

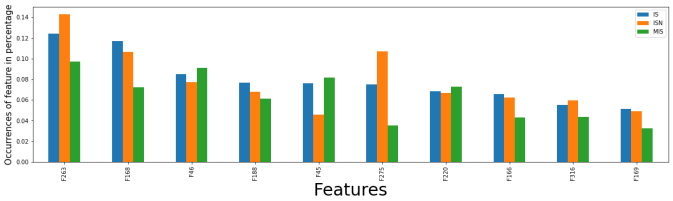

Fig. 17: Feature occurrence for top three features with the most importance in percentage according to the results of IG. These results are for the Volvo dataset for GRU model, and depend on the baselines.

integrated gradients are sensitive to noise, and therefore, the results can vary depending on the dataset. Even a dataset with noise can get a similar result as the theory depending on the model. To strengthen this theory, we applied integrated gradients with three different gradient baselines on SVR and SVM models with non-sequential data, which gave similar results as it did for the MLP model.

To answer the questions 1. "*How can the baseline for integrated gradients be chosen for tabular data in predictive maintenance?*" and 2. *"How does the baseline affect the outcome of different models?"*, we have applied IG on several datasets and models, with different gradient baselines. However, since IG is a local XAI method, we implemented an aggregated version of the method to get a larger view of how the model makes its predictions. This makes it easier to see how the features impact the whole model, and not only one prediction, which is very useful in predictive maintenance. We can see from the experiments and conclusions that the gradient baseline has a more significant impact on the results of IG on tabular and non-sequential data on models such as MLP and SVM. In contrast, the impact of the gradient baselines decreases for sequential models such as RNN, LSTM, and GRU. These results could imply that the IG method is more suited for sequential data. Furthermore, we observe that applying IG to a deep learning model provides knowledge on the importance of the features differently depending on how robust the data is when using data with time series. With noisy data, such as the Volvo dataset, the IG has a more challenging time making clear conclusions. This is not anything new, considering we know that IG is sensitive to noise; however, it is now more evident that this also applies to data with time series and not only for images.

We would have liked to discuss the results with domain experts, which is something we will bring for future work. We would also like to see a combination of integrated gradients with another gradient-based method as a future work within the area of XAI and predictive maintenance for time-series data.

ACKNOWLEDGMENT

This work was supported by research grants from KK-Foundation, Sweden.

REFERENCES

[1] B. Shukla, I.-S. Fan, and I. Jennions, "Opportunities for explainable artificial intelligence in aerospace predictive maintenance," in *PHM Society European Conference*, vol. 5, pp. 11–11, 2020.

[2] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[3] M. Van Lent, W. Fisher, and M. Mancuso, "An explainable artificial intelligence system for small-unit tactical behavior," in *Proceedings of the national conference on artificial intelligence*, pp. 900–907, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004.

[4] M. Rahat, S. Pashami, S. Nowaczyk, and Z. Kharazian, "Modeling turbocharger failures using markov process for predictive maintenance," in *30th European Safety and Reliability Conference (ESREL2020) & 15th Probabilistic Safety Assessment and Management Conference (PSAM15), Venice, Italy, 1-5 November, 2020*, European Safety and Reliability Association, 2020.

[5] V. Revanur, A. Ayibiowu, M. Rahat, and R. Khoshkangini, "Embeddings based parallel stacked autoencoder approach for dimensionality reduction and predictive maintenance of vehicles," in *IoT Streams for Data-Driven Predictive Maintenance and IoT, Edge, and Mobile for Embedded Machine Learning*, pp. 127–141, Springer, 2020.

[6] M. G. Altarabichi and et al., "Stacking ensembles of heterogenous classifiers for fault detection in evolving environments," in *30th European Safety and Reliability Conference, ESREL 2020 and 15th Probabilistic Safety Assessment and Management Conference, PSAM15 2020, Venice, Italy, 1-5 November, 2020*, pp. 1068–1068, Research Publishing, 2020.

[7] M. Rahat, P. S. Mashhadi, S. Nowaczyk, T. Rognvaldsson, A. Taheri, and A. Abbasi, "Domain adaptation in predicting turbocharger failures using vehicle's sensor measurements," in *PHM Society European Conference*, vol. 7, pp. 432–439, 2022.

[8] A. Saxena and K. Goebel, "Turbofan engine degradation simulation data set," *NASA Ames Prognostics Data Repository*, pp. 1551–3203, 2008.

 ISBN: ISBNFILL

[9] C. Reinartz, M. Kulahci, and O. Ravn, "An extended tennessee eastman simulation dataset for fault-detection and decision support systems," *Computers & Chemical Engineering*, vol. 149, p. 107281, 2021.
[10] C. C. Reinartz, M. Kulahci, and O. Ravn, "Tennessee Eastman Reference Data for Fault-Detection and Decision Support Systems," 2021.
[11] S. Matzka, "Explainable artificial intelligence for predictive maintenance applications," in *2020 Third International Conference on Artificial Intelligence for Industries (AI4I)*, pp. 69–74, 2020.
[12] A. Torcianti and S. Matzka, "Explainable artificial intelligence for predictive maintenance applications using a local surrogate model," in *2021 4th International Conference on Artificial Intelligence for Industries (AI4I)*, pp. 86–88, 2021.
[13] O. Serradilla, E. Zugasti, J. Ramirez de Okariz, J. Rodriguez, and U. Zurutuza, "Adaptable and explainable predictive maintenance: Semi-supervised deep learning for anomaly detection and diagnosis in press machine data," *Applied Sciences*, vol. 11, no. 16, 2021.
[14] A. P. Hermawan, D.-S. Kim, and J.-M. Lee, "Predictive maintenance of aircraft engine using deep learning technique," in *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 1296–1298, 2020.
[15] T. Haj Mohamad, A. Abbasi, E. Kim, and C. Nataraj, "Application of deep cnn-lstm network to gear fault diagnostics," in *2021 IEEE International Conference on Prognostics and Health Management (ICPHM)*, pp. 1–6, 2021.
[16] P. Sturmfels, S. Lundberg, and S.-I. Lee, "Visualizing the impact of feature attribution baselines," *Distill*, 2020. https://distill.pub/2020/attribution-baselines.
[17] Y. B. M.R. Gazarra, D.Singh, "Enhanced integrated gradients: improving interpretability of deep learning models using splicing codes as a case study," *Genome Biology*, vol. 21, no. 149, 2020.

 ISBN: ISBNFILL