Explaining Credit Card Fraud Decisions in ML: An Analysis of XAI Methods*

Ciaran Finnegan[1][0009−0008−9620−5460] and Dr. Bujar Raufi[1][1111−2222−3333−4444]

[1]Dublin Institute of Technology, Ireland

Abstract. Data Scientists working in the domain of commercial financial crime prevention software are aware of the various Explainable Artificial Intelligence (XAI) techniques that can be applied in the domain of credit card fraud, but relatively little published research on the comparative benefits of such approaches is still being done in this context. Furthermore, a significant volume of research is focused on the human interpretation of the XAI output, regardless of whether the subject is fraud detection or health care prediction. Human surveys are costly to implement and can be susceptible to bias and/or lack of domain knowledge by the participant. The focus of this paper is to look at an automated and statistical comparison of established XAI methods for credit card fraud and assess whether there is a quantitative difference between them in terms of general performance. Such an analysis could provide guidance for future product roadmaps in the commercial online fraud prevention space.

Keywords: Explainable Artificial Intelligence · Credit Card Fraud Detection · XAI Statistical Comparison.

1 Introduction

1.1 Interpretability in Credit Card Fraud Classification

The need for ever more sophisticated Machine Learning techniques to tackle the problem of credit card fraud has been well established by academic observers such as (Dal Pozzolo et al., 2014). The research of (Sharma & Bathla, 2020) and (Batageri & Kumar, 2021) is an example of work in this field to improve fraud detection rates through ever more sophisticated neural network algorithms. However, many researchers highlight the parallel challenge that these *'black box'* models need to be held accountable for the individual fraud classifications they make (T.Y.Wu & Y.T.Wang, 2021).

(Ignatiev, 2020) focuses on the need for Explainable Artificial Intelligence (XAI) to be *trustable*, while (Carvalho, Pereira, & Cardoso, 2019) are more emphatic about the legal demands of the European Union that all automated decision-making about citizens be *transparent*.

* Supported by Dublin Institute of Technology

This article will focus on whether an objective rating can be given to different XAI methods in terms of explaining the reason for a given credit card fraud classification. To further narrow the field of interest, the paper will propose a series of metrics to rate the performance of four state-of-the-art XAI methods; SHAP, LIME, ANCHORS, and DiCE on an industry credit card fraud dataset, as applied to the classification of individual credit card transactions.

1.2 Research Question

"To what extent can we quantify the quality of contemporary machine learning interpretability techniques, providing local, model-agnostic, and post-hoc explanations, in the classification of credit card fraud transactions by a 'black box' Neural Network ML model?"

The question will focus on a quantitative comparison of explanations produced by different XAI techniques on specific (local) NN model predictions.

1.3 Research Problem

The research problem can be described as the means to produce an objective assessment of state-of-the-art ML explainers, as applied to credit card fraud detection. The intention is to compare a set of common XAI techniques and to find insights into the relative strengths of each one. The focus of the experiment is on the application of SHAP, LIME, ANCHORS, and DiCE interpretability methods upon a Neural Network model trained on a commercial dataset containing credit card transactions, which are labelled *'fraud'* or *'non-fraud'*.

2 Background

2.1 Key Themes in Current Research

How to Measure the Effectiveness of an Explanation? No Obvious Consensus In their research experiments with the LIME algorithm, (Ribeiro, Singh, & Guestrin, 2016) describe how users can have a trust problem with NN ML models because they are effectively *'black-boxes'* from which it is very difficult to interpret why a given classification has been derived. This is a theme echoed in the introduction of many research papers, such as (ElShawi, Sherif, Al-Mallah, & Sakr, 2020), (Honegger, 2018), and (Sinanc, Demirezen, & Sağıroğlu, 2021). Although universal frameworks to interpret model predictions have been proposed (Lundberg & Lee, 2017) there is still no unanimity seen in research to date on what constitutes an objectively *'good'* explanation of a prediction. The gap remains; How exactly does a researcher measure and display *'explainability'* in Explainable Artificial Intelligence (XAI) research?

To further emphasise this gap in contemporary research, (Adadi & Berrada, 2018) claimed that *"Technically, there is no standard and generally accepted definition of explainable AI"* (p. 141). Therefore, there is no well-established output framework for explaining credit card fraud classification (Vilone & Longo, 2021).

This article proposes to build on some of the objective research on scoring predictions generated by four established interpretability methods.

Human Assessment vs Automated Benchmarks Research by (Jacob et al., 2021) makes the point about evaluating XAI output that *"...while a user study may be the best way to evaluate the usefulness of explanations, it is not always available and may come at a high cost."*. It is also desirable for humans participating in XAI surveys to have some degree of domain knowledge, but fraud detection explainer experiments by (Jesus et al., 2021) showed that this can still be subject to user bias.

Examples of XAI research where the reliance on human assessment of explanations is less commonplace can be seem in the domain of healthcare, through research by (Marcilio & Eler, 2020) and (Lakkaraju, Bach, & Leskovec, 2016). These experiments produce clear objective recommendations in line with the work of (ElShawi et al., 2020).

This thesis will follow in the steps of earlier research that only use nonhuman programmatic experiments with quantifiable metrics (Darias, Caro-Martínez, Díaz-Agudo, & Recio-Garcia, 2022) and tests for statistical significance (Evans, Xue, & Zhang, 2019).

2.2 State of the Art Approaches for Local Interpretability

This section of the document describes research conducted on local interpretability techniques that formed the basis of the experiments in this dissertation.

SHAP SHAP stands for **SH**apley **A**dditive ex**P**lanations (Lundberg & Lee, 2017) and can be described as a unified framework for interpreting predictions. SHAP is a method derived from cooperative game theory, and SHAP values are used extensively to present an understanding of how the features in a dataset are related to the model prediction output.

LIME LIME stands for **L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations (Ribeiro et al., 2016) and is also a popular choice for interpreting decisions made by black box models. The core concept of LIME is that it aims to understand the features that influence the prediction of a given black-box model around a single instance of interest.

ANCHOR ANCHORS was also developed by Marco Ribeiro (Ribeiro, Singh, & Guestrin, 2018) and is, again, a model-agnostic explanation approach based on if-then rules that are called *'anchors'*. These *'anchors'* are a set of feature conditions that act as high precision explainers created using reinforcement learning methods.

DiCE DiCE (Diverse Counterfactual Explanations) (Mothilal & Tan, 2020) is an XAI method developed to provide information on the decisions of the machine learning model by generating counterfactual explanations. In essence, a counterfactual explanation describes a minimal set of changes required to alter the model's prediction for a particular instance.

3 Methodologies

3.1 Research Objectives and Experimental Activities

The study will run an iteration of the eight following research steps to compile a table of metric results for each explainer method.

These steps will build a statistical comparative analysis of the performance of each technique;

1. Train, test, and evaluate a credit card fraud NN detection model.
2. Generate explanation(s) for each method based on a small subset of test data. Produce visual validation that the explainer output is meaningful.
3. If necessary, refine the model-building process to improve the quality of the explanations.
4. Break out the test data into equal blocks of feature instances, with associated fraud labels, and generate explanations for the instances in each block. Convert the output to numerical values, as appropriate.
5. Submit the XAI output data to a separate Python function to generate a value from each experiment metric (See 3.2 for further elaboration).
6. Take the average metric score for each metric for each block and use these values as input for a statistical comparison.
7. Review to determine whether any distortion occurred during the conversion of the XAI method outputs to numerical values. Correct as appropriate.
8. Conduct a comparative statistical analysis for each XAI method and determine if any significant performance difference can be proven.

The research focus is on explanations for fraud classification of individual transaction records; hence these experiments only consider local, post-hoc results.

Assessing the Explanations from Each XAI Method The records in the test data block will generate a table of numerical outputs against the following metrics (elaborated in Section 3.2 of this article);

1. Identity
2. Stability
3. Separability
4. Similarity
5. Computational Efficiency

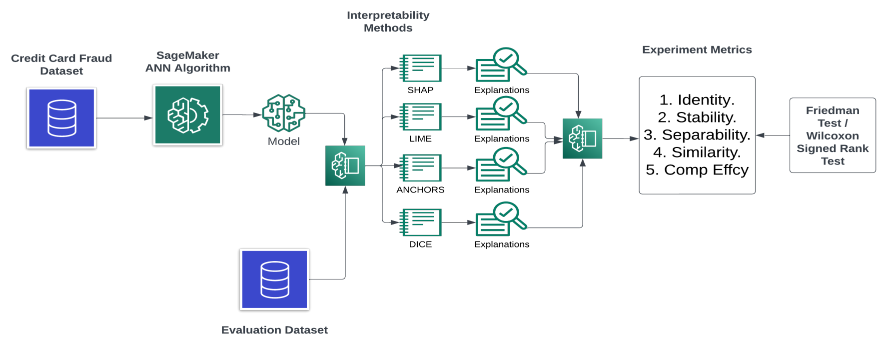

Fig. 1. Overview of experiment design

Figure 1 shows the diagrammatic view of the experiment design to compare explainability methods.

3.2 Experiment Design: Evaluation of XAI metrics and Statistical Analysis

The explainability metrics proposed below extend the framework comparison research conducted by (ElShawi et al., 2020), but transfer the domain from healthcare analysis to credit card fraud detection.

For this article, the metrics in the research referenced above have been adapted and extended to measure the following XAI characteristics;

1. Identity. A measure of how many identical instances have identical explanations. For every two instances in the testing data if the distance between features is equal to zero, then the distance between the explanations should be equal to zero.
2. Stability. Instances belonging to the same class have comparable explanations. K-means clustering is applied to the explanations for each instance in the test data. Measure the number of explanations in both clusters (fraud/non-fraud) that match the predicted class.
3. Separability. Dissimilar instances must have dissimilar explanations. Take a subset of test data and determine for each individual instance the number of duplicate explanations in the entire subset, if any.
4. Similarity. This metric captures the assumption that the more similar the instances to be explained, the closer their explanation should be (and vice versa). Cluster test data instances into Fraud/non-Fraud clusters. Normalise explanations and calculate Euclidean distances between instances in both clusters. Smaller mean pairwise distance = better explainability framework metric.
5. Computational Efficiency. Average time taken, in seconds, by the interpretability framework to generate output.

This will be a deductive approach to test the assumption that one particular interpretability framework can be shown, through statistical significance testing on the numerical outputs of each experiment, to generate the best local explanations for a credit card fraud classification result.

A Friedman test will be run to determine if there is evidence that there is a statistical difference in performance between SHAP, LIME, ANCHORS, and DiCE in terms of explaining local credit card fraud classification results.

4 Results

4.1 SHAP: XAI Experiment Results

Verification of SHAP Explainer Each of the four experiments on machine learning interpretability techniques in this thesis began with a verification that the explainers actually produce meaningful output.

This thesis is following the objective of defining a purely statistical approach to evaluating XAI methods at scale, but a visual (manual) inspection is carried out first to confirm the validity of each XAI method.

This check of the SHAP values involves a summary plot generated from the first 25 rows in the test data set. The SHAP explainer is created based on the NN Credit Card predictive model built in a previous step, along with a selection of training data. The *shap.summary_plot()* function is used to display the top 20 features that influence the classification of *non-fraud*.

SHAP Output Results from Metric Scoring The SHAP explainer experiments produced the following table of results. Each sample row in the table in figure 2 represents the metrics calculated in a sequential block of the test data set.

Sample Number	XAI_Identity	XAI_Stability	XAI_Seperability	XAI_Similairity	Comp_Efficiency
1	40.0000	87.6923	96.9231	0.2746	384.51
2	46.1538	76.9231	96.9231	0.2320	386.19
3	49.2308	67.6923	90.7692	0.2071	384.94
4	36.9231	75.3846	93.8462	0.2637	386.69
5	47.6923	83.0769	100.0000	0.2800	384.34
6	44.6154	27.6923	93.8462	0.2144	391.56
7	36.9231	75.3846	100.0000	0.3984	386.00
8	38.4615	13.8462	100.0000	0.2623	390.81
9	46.1538	70.7692	96.9231	0.3450	393.99
10	43.0769	33.8462	96.9231	0.3944	390.54
11	40.0000	84.6154	100.0000	0.2189	397.76
12	41.5385	32.3077	100.0000	0.3294	398.10
13	38.4615	70.7692	100.0000	0.2444	399.48
14	27.6923	80.0000	100.0000	0.2967	396.85
15	32.3077	75.3846	95.3846	0.3791	394.65
16	35.3846	76.9231	95.3846	0.3148	390.70
17	43.0769	29.2308	100.0000	0.2203	392.06
18	53.8462	29.2308	96.9231	0.2233	386.87
19	43.0769	23.0769	93.8462	0.2148	381.84
20	41.5385	46.1538	100.0000	0.3053	387.78

Fig. 2. SHAP XAI Experiment: Metrics Scores

Explaining the XAI Metric Scorecard (for All XAI Methods)

– The Sample Number identifies the individual data *chunk* extracted from the test dataset.
– *XAI Identity* is the separate score obtained from each data *chunk*. This is a score that can range from zero to 100.
– *XAI Stability* is also a score in the range of zero to 100 for each data *chunk*.
– *XAI Seperability* is another score of 0 - 100.
– *XAI Similarity* is a Euclidean measure of the average distance between points scored for this metric for each data *chunk*.
– *Computational Efficiency* is the time taken in seconds for the XAI method to actually generate explanations for each data *chunk*.

The mean of each set of column values is used as input to the statistical analysis described in Section 4.5.

4.2 LIME XAI Experiments: Results

Verification of LIME Explainer The Python LIME (Local Interpretable Model-agnostic Explanations) library generates explanations for individual predictions of any classifier or regressor, by approximating the model locally around each data point.

Following the established steps for these XAI metrics experiments, a random instance was selected from the test data to verify the output of the *lime_tabular()* Python function.

LIME Output Results from Metric Scoring Each sample row in the table in Figure 3 below represents the LIME metrics calculated on a sequential block from the test dataset.

Sample Number	XAI_Identity	XAI_Stability	XAI_Seperability	XAI_Similairity	Comp_Efficiency
1	10.8214	26.2323	100.0000	0.6541	754.96
2	3.0769	78.4615	90.7692	0.5767	766.21
3	3.0769	60.0000	100.0000	0.5457	726.49
4	4.6154	33.8462	93.8462	0.6235	749.68
5	12.3077	40.0000	100.0000	0.6593	740.32
6	1.5385	24.6154	96.9231	0.5396	724.98
7	3.0769	69.2308	90.7692	0.5989	733.20
8	4.6154	49.2308	100.0000	0.6556	754.26
9	10.7692	26.1538	100.0000	0.6376	744.94
10	4.6154	35.3846	100.0000	0.6671	730.45
11	3.0769	66.1538	100.0000	0.5025	732.48
12	3.0769	63.0769	93.8462	0.6177	731.88
13	3.0769	81.5385	93.8462	0.5432	742.28
14	1.5385	67.6923	100.0000	0.5480	728.85
15	3.0769	26.1538	93.8462	0.6075	728.72
16	0.0000	73.8462	100.0000	0.6395	727.66
17	6.1538	40.0000	96.9231	0.5484	730.83
18	4.6154	32.3077	90.7692	0.5875	739.91
19	6.1538	36.9231	96.9231	0.4402	745.11
20	3.0769	64.6154	96.9231	0.6025	729.68

Fig. 3. LIME XAI Experiment: Metrics Scores

4.3 ANCHORS XAI Experiments: Results

Verification of Anchor Explainer In the context of Explainable AI (XAI), the Python ANCHOR library generates feature-specific explanations for individual instances in a test dataset by identifying minimal sets of conditions, or 'anchors', that are sufficient to ensure the same prediction for similar instances.

Again, random instances were selected from the test data to verify the output of the *anchor_tabular()* Python function.

Anchor Output Results from Metric Scoring The ANCHORS explainer experiments produced the following table of results. Each sample row in the table represented in figure 4 represents the metrics calculated on a sequential block from the test dataset.

Sample Number	XAI_Identity	XAI_Stability	XAI_Seperability	XAI_Similairity	Comp_Efficiency
1	15.6250	82.8125	15.6250	1.5377	2901.37
2	12.5000	60.9375	20.3125	1.1226	2783.96
3	6.2500	60.9375	25.0000	1.7104	2799.25
4	9.3750	32.8125	20.3125	1.4956	2803.50
5	6.2500	20.3125	18.7500	1.7493	2726.46
6	12.5000	43.7500	17.1875	1.0614	3072.32
7	4.6875	29.6875	25.0000	2.1536	2903.02
8	7.8125	60.9375	32.8125	1.2250	2713.01
9	1.5625	26.5625	23.4375	1.7183	2796.82
10	9.3750	53.1250	17.1875	2.0590	3005.47
11	9.3750	65.6250	15.6250	1.0507	2647.00
12	12.5000	45.3125	20.3125	1.7524	3148.17
13	18.7500	60.9375	25.0000	1.6294	2722.98
14	9.3750	73.4375	20.3125	1.8468	2672.22
15	3.1250	28.1250	18.7500	2.0189	2627.63
16	14.0625	43.7500	17.1875	1.5749	2791.43
17	7.8125	40.6250	25.0000	1.6403	2900.18
18	12.5000	43.7500	32.8125	2.3355	2914.33
19	10.9375	60.9375	23.4375	1.0377	2735.07
20	9.3750	64.0625	17.1875	1.8237	2994.62

Fig. 4. ANCHOR XAI Experiment: Metrics Scores

4.4 DiCE XAI Experiments: Results

Verification of DiCE Explainer The Python DiCE library generates counterfactual explanations for individual instances in a test dataset, focussing on identifying minimal changes to the feature values that would alter the model's prediction.

DiCE Output Results from Metric Scoring The DiCE explainer experiments produced the following table of results. Each sample row in Table 5 represents the metrics calculated in a sequential block from the test dataset.

Sample Number	XAI_Identity	XAI_Stability	XAI_Seperability	XAI_Similairity	Comp_Efficiency
1	6.1538	58.4615	35.3846	13.1316	76.47
2	18.4615	46.1538	36.9231	16.2164	77.33
3	9.2308	56.9231	41.5385	13.8937	122.52
4	24.6154	55.3846	40.0000	14.9150	75.33
5	7.6923	55.3846	49.2308	13.9490	121.45
6	20.0000	43.0769	49.2308	19.9329	79.69
7	15.3846	60.0000	52.3077	19.7095	131.22
8	13.8462	55.3846	44.6154	12.3597	75.06
9	18.4615	47.6923	49.2308	14.6446	76.56
10	6.1538	50.7692	46.1538	18.9109	84.86
11	12.3077	64.6154	53.8462	15.0291	83.61
12	10.7692	66.1538	49.2308	15.7752	79.67
13	16.9231	58.4615	40.0000	19.0540	89.36
14	7.6923	73.8462	53.8462	14.8314	75.93
15	9.2308	64.6154	50.7692	16.7801	82.30
16	9.2308	56.9231	49.2308	15.1260	76.56
17	7.6923	61.5385	40.0000	16.5128	76.23
18	13.8462	58.4615	58.4615	16.9805	76.22
19	24.6154	44.6154	44.6154	13.7436	80.85
20	4.6154	60.0000	55.3846	18.4578	78.25

Fig. 5. DiCE XAI Experiment: Metrics Scores

4.5 Aggregate XAI Experiment Results

The final output generated by the XAI metrics experiments produces the following matrix of mean values, as displayed in Table 1.

Table 1. Final Table of XAI Metrics Results

	SHAP	**LIME**	**ANCHORS**	**DiCE**
Identity	41.308	4.618	9.688	12.846
Stability	58.000	49.773	49.922	56.923
Separability	97.385	96.769	21.563	47.000
Similarity	0.281	0.590	1.627	15.998
Computational Efficiency	390.282	738.145	2832.941	85.974

4.6 Friedman Test Analysis

Tabular View of Friedman Results Given that the custom XAI metrics of the research experiments comprises different measures applied to each XAI method, the Friedman test effectively evaluates the NULL hypothesis that these methods ***do not*** differ significantly in their performance.

Using the data in Table 1 as the input to a Friedman test, the following statistics were generated;

Table 2. Friedman Test Statistics

Statistic	1.5600
P-Value	0.6685

The Friedman test statistic is approximately 1.56 with a p-value of approximately 1.56.

This result indicates that there is no statistically significant difference between the four XAI methods (SHAP, LIME, ANCHORS, and DiCE) when comparing results on our chosen credit card fraud dataset. This analysis is based on the numbers captured for the XAI Metrics in the earlier experiments.

5 Conclusion

5.1 Summary

The experiments in this thesis showed that, after training a Neural Network model on a credit card fraud dataset, it was **not** possible to distinguish between the merits of the SHAP, LIME, ANCHORS, and DiCE interpretability methods.

Validity of the Statistical Analysis of the XAI Techniques? The experiments carried out in this research established that repeatable statistical analysis was possible and that comparisons could be drawn between different XAI techniques.

During the execution of the XAI metrics experiments and the analysis of the results, the following types of observations emerged about this type of statistical analysis.

- The output from LIME, Anchors, and DiCE is immediately understandable by a human reader. It could be argued that this is the strength of their output and that these techniques have not been built for statistical analysis. However, this was one of the stated objectives of this research, to tailor this type of interpretability output for a statistical comparison analysis.
- The DiCE algorithm was ineffective in producing counterfactual explanations until an increased number of continuous features, with a wider range of values, were added to the model creation process. The necessity to 'pad' the feature list with additional continuous predictors raises questions around the suitability of an XAI technique such as DiCE counterfactuals in future experiments.
- The computational overhead to generate the Anchor explanations almost derailed the execution of the experiment itself. It could be theorised that additional preprocessing of the credit card dataset, or a refinement of the *Anchor_Tabular* algorithm, might mitigate this complexity. However, such an analysis is beyond the current scope of this thesis.

The conclusion drawn from the points above, which stem from the observations on both the execution and results of the experiments in this thesis, is that the researcher must be cognisant of **how** the XAI technique manages the characteristics of the source data. The statistical analysis in this article is a useful tool, and there is insight to be gained, but the value of the metric scores must be considered in conjunction with how the individual experiment handled the source data.

Thus, the conclusion tends to support that this analysis does provide insight, but the interaction of data and interpretability technique have to be carefully considered.

5.2 Recommendations for Future Research

Broaden Range of XAI Techniques The experiments in this paper were limited to four XAI techniques, but fraud classification would also be suitable for interpretability processes such as **LORE**, **ILIME**, **MAPLE**. An obvious evolution is to extend the breath of explainers, introduce new datasets, and increase the matrix of metrics input into the Friedman/Wilcoxon-Paired tests.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, *6*(52), 138–160. doi: https://doi.org/10.1109/access.2018.2870052

Batageri, A., & Kumar, S. (2021). Credit card fraud detection using artificial neural network. *Global Transitions Proceedings*, *2*(1), 35–41. doi: https://doi.org/10.1016/j.gltp.2021.01.006

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019, Jul). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8). doi: https://doi.org/10.3390/electronics8080832

Dal Pozzolo, A., et al. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications*, *41*(10), 4915–4928. doi: https://doi.org/10.1016/j.eswa.2014.02.026

Darias, J. M., Caro-Martínez, M., Díaz-Agudo, B., & Recio-Garcia, J. A. (2022, Aug). Using case-based reasoning for capturing expert knowledge on explanation methods. *Case-Based Reasoning Research and Development*, *13405*, 3–17. doi: https://doi.org/10.1007/978-3-031-14923-8_1

ElShawi, R., Sherif, Y., Al-Mallah, M., & Sakr, S. (2020, Aug). Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Computational Intelligence*, *37*(4), 1633–1650. doi: https://doi.org/10.1111/coin.12410

Evans, B. P., Xue, B., & Zhang, M. (2019, Jul). What's inside the black-box? *Proceedings of the Genetic and Evolutionary Computation Conference*. doi: https://doi.org/10.1145/3321707.3321726

Honegger, M. (2018, Aug). *Shedding light on black box machine learning algorithms: Development of an axiomatic framework to assess the quality of methods that explain individual predictions.* Karlsruhe Institute of Technology. Retrieved from https://arxiv.org/abs/1808.05054v1

Ignatiev, A. (2020, Jul). Towards trustable explainable ai. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 5154–5158. doi: https://doi.org/10.24963/ijcai.2020/726

Jacob, V., Song, F., Stiegler, A., Rad, B., Diao, Y., & Tatbul, N. (2021). Exathlon: A benchmark for explainable anomaly detection over time series. *Proceedings of the VLDB Endowment*, *14*(11), 2613–2626. doi: https://doi.org/10.14778/3476249.3476307

Jesus, S., Belém, C., Balayan, V., Bento, J., Saleiro, P., Bizarro, P., & Gama, J. (2021, Mar). How can i choose an explainer? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. doi: https://doi.org/10.1145/3442188.3445941

Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016, Aug). Interpretable decision sets. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1675–1684. doi: https://doi.org/10.1145/2939672.2939874

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems 30 (nips 2017)* (Vol. 30). NeurIPS Proceedings.

Marcilio, W. E., & Eler, D. M. (2020, Nov). From explanations to feature selection: Assessing shap values as feature selection mechanism. *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 340–347. doi: https://doi.org/10.1109/sibgrapi51738.2020.00053

Mothilal, S. A., R. K., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. doi: https://doi.org/10.48550/arXiv.1806.08049

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, Aug). "why should i trust you?" explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. doi: https://doi.org/10.1145/2939672.2939778

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018, Feb). Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). doi: https://doi.org/10.1609/aaai.v32i1.11491

Sharma, A., & Bathla, N. (2020, Aug). *Review on credit card fraud detection and classification by Machine Learning and Data Mining approaches*, *6*(4), 687–692.

Sinanc, D., Demirezen, U., & Sağıroğlu, (2021). Explainable credit card fraud detection with image conversion. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, *10*(1), 63–76. doi: https://doi.org/10.14201/adcaij20211016376

T.Y.Wu, & Y.T.Wang. (2021). Locally interpretable one-class anomaly detection for credit card fraud detection. *2021 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*. doi: https://doi.org/10.1109/taai54685.2021.00014

Vilone, G., & Longo, L. (2021, May). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, *76*, 89–106. doi: https://doi.org/10.1016/j.inffus.2021.05.009