Assessment of Explainable AI Approaches for Interpreting Black-Box Models in Credit Card Fraud Detection: A Comparative Study of SHAP, LIME, ANCHORS, and DICE Methods *

Bujar Raufi[1[1111−2222−3333−4444]], Ciaran Finnegan[1[0009−0008−9620−5460]], and Luca Longo[1[000−0002−2718−5426]]

Technological University Dublin, Dublin, Ireland
bujar.raufi,luca.longo@tudublin.ie
d21124026@mytudublin.ie
https://www.tudublin.ie/explore/faculties-and-schools/computing-digital-data/computer-science/

Abstract. Financial institutions heavily rely on advanced Machine Learning algorithms to screen transactions. However, they face increasing pressure from regulators and the public to ensure AI accountability and transparency, particularly in credit card fraud detection. While ML technology has effectively detected fraudulent activity, the opacity of Artificial Neural Networks (ANN) can make it challenging to explain decisions. This has prompted a recent push for more explainable fraud prevention tools. Although vendors claim to improve detection rates, integrating explanation data is still in its early stages. Data scientists recognize the potential of Explainable AI (XAI) techniques in fraud prevention, but comparative research on their effectiveness is lacking.

This paper aims to advance the comparative research on credit card fraud detection by statistically evaluating established XAI methods. The goal is to explain and validate the fraud detection black-box machine learning model, where the baseline model used for explanation is an ANN trained with a large dataset of 25,128 instances. Four explainability methods (SHAP, LIME, ANCHORS, and DiCE) are utilized, and the same test set is used to generate an explanation across all four methods. Analysis through the Friedman test indicates a statistical significance of the SHAP, ANCHORS, and DiCE results, validated with interpretability and reliability aspects of explanations such as identity, stability, separability, similarity, and computational complexity. The results indicated that SHAP, LIME, and ANCHORS methods exhibit better model interpretability regarding stability, separability, and similarity.

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

* Supported by Technological University Dublin

1 Introduction

Credit card fraud costs the Financial Services industry billions of Euros in losses each year[19]. The need for ever more sophisticated Machine Learning techniques to tackle this problem has been well established in the academic community [5], [26]. Research outlined in [25] and [3] are examples 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 that they make [29]. The need for a more trustable Explainable Artificial Intelligence (XAI) as well as the transparency of automated decision-making is gaining momentum as it becomes part of European legal demands [11], [4].

Producing an objective assessment of state-of-the-art ML explainers, as applied to credit card fraud detection, is essential. The intention is to compare standard XAI techniques and look for insights into each one's relative strengths. The experiment focus is on the application of SHAP, LIME, ANCHORS, and DiCE interpretability methods upon a A Neural Network model was trained on a commercial dataset containing credit card transactions, labelled into two classes indicating regular and fraud activities.

This paper focuses on the explainability of machine learning-driven software used by the Financial Services industry and whether different XAI methods can be objectively rated to explain a given credit card fraud classification behind artificial neural network (ANN) models. To further narrow the field of interest, the paper proposes 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. Specifically, the scope of experiments is on explanations for individual ('local') transactions and only considers interpretability techniques that are agnostic about the type of the detection model.

The paper addresses the following research question:

"To what extent can we quantify the quality of machine learning explainability techniques, providing local, model-agnostic, and post-hoc explanations, in classifying credit card fraud transactions by a 'black box' Artificial Neural Network (ANN) Machine Learning Model?"

The question will quantitatively compare explanations produced by different XAI techniques on specific (local) ANN model predictions.

The rest of the paper is organized as follows: section 2 outlines the related work on the model explainability research in the context of financial fraud detection, section 3 provides the experiment design that tackles the aforementioned research question; section 4 elaborates the research results and finding and 5 concludes the paper together with future work.

2 Related Work

The use of Machine Learning (ML) models in detecting credit card fraud has evolved significantly, focusing on increasingly sophisticated neural network architectures [26]. However, there are still challenges that need to be addressed, particularly when it comes to measuring the effectiveness of explanations provided by neural network models. Additionally, there is a need for better ways to assess model outputs and account for the computational efficiency of such outputs at runtime.

In terms of measuring the effectiveness of explanations in interpretability of ML models, particularly neural networks, which are often regarded as "black boxes" due to the difficulty in understanding their decision-making process, represents a challenge [22]. This challenge of interpretability is echoed in various studies across XAI community, underscoring the lack of a definitive approach to establishing trustworthiness in eXplainable Artificial Intelligence (XAI) systems within the domain of credit card fraud detection [7], [10], [27].

Despite efforts to develop universal frameworks for interpreting Machine Learning (ML) model predictions [15], there remains a notable absence of consensus on what constitutes a satisfactory explanation for a prediction. This gap is further emphasized by the absence of a standard definition for explainable AI [1]. Similarly, there is an evident lack of consensus in the research community regarding the nature of explanations and the essential properties required to make them understandable to end-users [30].

Evaluating the usefulness of explanations through direct human assessment is valuable, but it may not always be available and can come at a high cost [12]. It is also desirable for humans taking part in XAI explanations to have some degree of domain knowledge. However, fraud detection explainer experiments showed that this can still be subject to user bias [13]. Research into explanations for ML fraud classification often follows a more subjective survey style of experimentation involving the augmentation of human-based processes with model explainer outputs. Occasionally, human bias can adversely impact the reliability of the interpretation of the ML-generated model explanations [14]. This paper will focus on generating model explanations without direct human involvement to avoid bias and provide more programmatic experiments with quantifiable metrics[6].

Much of the research on automated model explanation can be seen in anomaly detection framework for explainability [20] and XAI time-series results, both explained with SHAP and LIME explainer outputs[24]. Furthermore, a framework using local explanations generated with SHAP, LIME, and Counterfactual techniques can be seen in [16] and [9].

A significant body of research on XAI approaches is witnessed in using Deep Learning (DL) techniques such as Deep SHAP [18], [28], LIME for Deep Neural Network (DNN) experiments [21], and the use of counterfactuals in CNN model explainers [32].

2.1 Approaches to Local Interpretability in Explainable AI

This section outlines the research conducted on local interpretability techniques that served as the foundation for the experiments in this paper.

SHapley **A**dditive ex**P**lanations (SHAP) framework is a method that helps in interpreting predictions [15]. It is derived from cooperative game theory and is widely used to understand how the features in a dataset are related to the model prediction output. It provides globally consistent explanations and handles complex interactions well due to its strong theoretical grounding. However, it can be computationally complex, especially for large and higher dimensional datasets.

Local **I**nterpretable **M**odel-agnostic **E**xplanations (LIME) is a widely used method for interpreting decisions made by black box models [22]. The main idea behind LIME is to identify the features that have the most impact on predicting a specific instance of interest within the model. LIME is known for its simplicity and speed, which make it an excellent option for obtaining quick insights. However, it may not accurately capture complex relationships, especially in high-dimensional spaces.

ANCHORS is a model-agnostic explanation approach that provides interpretable rules for the model's predictions. Introduced by Ribeiro [23], it offers both local and global explanations. The strategy is based on if-then rules called *'anchors'*, where feature conditions act as high-precision explainers. These *'anchors'* are created using reinforcement learning methods. Although ANCHORS can generate easily understandable and precise rules, they may not provide a finer understanding of interactions available through other methods.

Diverse **C**ounterfactual **E**xplanations (DICE) is a powerful XAI method that effectively generates counterfactual explanations to provide insights into the decisions made by a machine learning model [17]. DICE offers actionable insights by identifying the minimum changes required to modify the model's prediction for a specific instance. It helps to address the common challenge of opaque decision-making in complex models. However, the method may struggle with high-dimensional data and complex models, limiting the diversity of counterfactuals generated.

Much of the research has been invested in providing a model explanation by comparing SHAP and LIME [33, 8]. It is worth noting that a direct comparison of SHAP, LIME ANCHORS and DICE to explain "black-box" ANN models in the context of credit card fraud detection models is generally lacking. Consequently, the paper identifies a significant void in contemporary research regarding establishing a comprehensive framework for explaining credit card fraud classifications made by "black-box" models [31]. In response to this gap, the paper proposes to build upon existing objective research by evaluating the performance of four established interpretability methods in scoring predictions.

3 Design Methodology

The research aim is to rank four selected interpretability frameworks (LIME, SHAP, Anchors, and DiCE), using predefined, custom-built comparison metrics, against the output of a Neural Network (NN) credit card fraud detection model and determine which one, if any, demonstrates the best overall performance. The design pipelines are done through dataset collection and preprocessing, hyperparameter tuning and model building, model evaluation, and model explanation metrics to achieve the above-mentioned.

3.1 Dataset and Preprocessing

The dataset utilized in this study was provided by a company and is associated with a product development cycle that occurred between 2014 and 2018. The data was collected in 2013 from several sources of credit card transactions within the United States and comprises 25,128 rows, with each row representing a credit card purchase. While this dataset was initially utilized for product testing and demonstration purposes, the product line it was connected with was terminated in 2019, and permission has been granted to access this now-defunct dataset.

Around 15% of the records in this research dataset are fraudulent, which is more balanced than typical credit card fraud datasets. However, a downsampling of the non-fraudulent records is applied to create an even classification split. Removal of a section of non-fraudulent records to simplify the process and avoid adding new synthetic data is performed. As a result, the remaining dataset have only 7,000 rows and a 50/50 breakdown of fraudulent and non-fraudulent records. The dataset comprises 380 initial features, and filtering methods is used to reduce the number of features. The final dataset consists of 15,000 rows and 40 features.

3.2 Model Building, Hyperparameter tuning and Training

A feed-forwards artificial neural network (ANN) is employed to classify fraudulent transactions. Before model building, a hyperparameter tuning is used to select the best model for training. The hyperparameter tuning used a random search method across six hyperparameters spread across ANN's input and hidden layers. Table 1 outlines the parameters, the search space and the selected value for the model.

Model training is done across 100 epochs using the Adam optimization algorithm with Binary Crossenrtropy loss function. A typical 80/20 split is employed for train and testing. A stop loss of patience three is employed during training to stop the model if no change in loss function is witnessed across epochs.

3.3 Model Evaluation Metrics

In various experiments conducted to detect fraud using Artificial Neural Networks (ANN), an accuracy score of $>= 0.85$ and an F1 score of $>= 0.85$ have

Table 1: Hyperparameter tuning search space and values

Parameter	Search Space	Selected Value
Input units	[32,...,512]	64
Input Dropout	[0.0,...,0.5]	0.25
No. hidden layers	[1, 3]	2
Hidden units	[32,...,512]	512 and 448
Hidden dropout	[0.0,...,0.5]	0.25
Learning rate	[0.01, 0.001, 0.0001]	0.01

been identified as ideal targets [27],[2]. While credit card fraud datasets are generally imbalanced, with very few instances of fraudulent behaviour, the dataset used in our experiment differs, as explained in section 3.1. Therefore, model's overall accuracy score is used to measure its performance in predicting fraud and non-fraud outcomes. The F1 score is a metric that combines precision and recall. Precision measures the correctly identified positive cases, while recall measures the true positive rate. In credit card fraud detection, both precision and recall are vital. High precision reduces inconvenience for customers, while high recall ensures financial security. The F1 score represents a harmonic man between recall and precision and focuses on the model's performance in predicting fraudulent transactions. The ROC curve is another metric used to evaluate binary classification models, such as fraud classification. Particular care of model evaluation is also given to errors made by the model during training, known as loss function. The model loss function is crucial in preventing the model from overfitting. Overfitting is remedied through a stop loss mechanism as elaborated in subsection 3.2.

3.4 Model Explanation Metrics

To assess the Explanations from Each XAI Method, a single predictive model for credit card fraud is built and stored for experiments to generate explainer output for each XAI method (SHAP, LIME, ANCHORS and DICE). The test data block will subsequently be sub-divided into 20 batches for use in the research experiments to generate a table of numerical outputs against the following metrics:

1. *Identity:* A measure of how much identical instances have identical explanations. For every two instances in the testing data, if the distance between features equals zero, the distance between the explanations should equal zero. A higher score indicates that the explanations provided align well with what the model is actually doing.
2. *Stability:* Instances belonging to the same class have comparable explanations. K-means clustering is applied to explain each instance in test data. Measure the number of explanations in both clusters (fraud/non-fraud) that match the predicted class. A higher stability score implies that the explana-

tions remain consistent across different scenarios, crucial for building trust in the model.

3. *Separability:* Dissimilar instances must have dissimilar explanations. Take a subset of test data and determine for each instance the number of duplicate explanations in the entire subset, if any. A higher separability score indicates that the explanations effectively highlight the features contributing to the distinction between different classes, making the model's decision-making process more transparent.
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 equalizes to a better explainability framework metric.
5. *Computational Efficiency:* Average time taken, in seconds, by the interpretability framework to output a set of explanations.

A metric such as '*Computational Efficiency*' could be considered unrelated to a measure of explainability. Still, this research contends that it is important to consider the feasibility/practicality of XAI methods. Computational time can be a bottleneck in generating explanations and may impact the commercial viability of an explainability process in a credit card fraud detection application.

The overall experiment design is outlined in figure 1. The study will execute

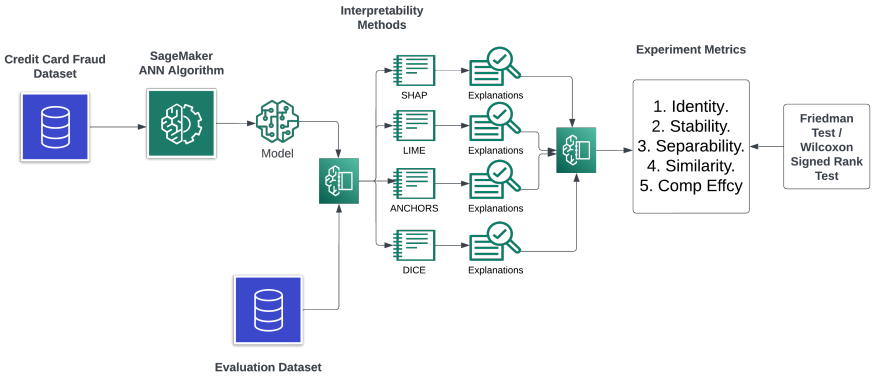


Fig. 1: Overview of experiment design

an iteration of the following eight research steps to compile a table of metric results for each explainer method. The steps involved in building a statistical comparative analysis of all four explainable AI techniques comprise the following:

1. Train, test, and evaluate a credit card fraud detection model built with ANN.

2. Generate explanation(s) for each method based on a single instance, or a minimal subset, taken from the test data. Use this output to display the explanation and validate the explainer visually.
3. Refine the credit card fraud model-building process if that improves the quality of the explanations without compromising on model performance.
4. Break out the test data into equal blocks of feature instances with associated fraud labels and generate explanations for each block's instances for each XAI method.
5. Generate the Identity, Stability, Separability, Similarity and Computation Efficiency metrics.
6. Take the average metric score from each metric for each block and use these values as the input to a statistical comparison. The reason for working with the average scores is that some XAI methods are computationally cumbersome, and processing the entire test data block at once was impractical.
7. Review the metric scores to determine if there could have been any distortion during converting XAI method outputs to numerical values. Correct as appropriate, and re-generate the XAI metric scores.
8. Conduct a comparative statistical analysis of the XAI metric score for each XAI method to determine if any significant difference in performance can be proven. This analysis will dictate the key results and observations of this thesis.

4 Results

4.1 Data Preprocessing results

The dataset of credit card transactions consisted of two labels in the target class: non-fraud (0) and fraud (1). A baseline preprocessing and model building was used to explain all four XAI methods. Table 3 illustrates processing activities done against the dataset regarding instances and features, removed features, outliers and the number of positive and negative train examples.

Table 2: Pre-processing activities against the dataset

Activity	**Value**
Instances	25128
Features	380
Rows Removed	297
Outliers removed	2340
Number of Features	64
Number Continuous Features	59
Number Categorical Features	5
Number Train Examples	5256
Number Positive Train Examples	2635
Number Negative Train Examples	2621

4.2 Evaluating the Predictive Fraud Model

This predictive credit card fraud model used in these experiments was required to demonstrate strong performance metrics to ensure that meaningful XAI data would be generated.

Figure2 illustrates the model accuracy evaluated with accuracy score, ROC areas under curve, recall, precision and f1-score metrics. The figure shows a balanced model performance in both classes, indicating consistent model building during training and good model generalizability.

Fig. 2: ANN Model's Class Accuracy

The figure shows a good performance of the model above 87% measured across all four evaluation metrics like accuracy score, ROC area under curve, recall, precision and f1-score.

4.3 Model Explanations and Comparisons

Concerning the model explainability evaluation using the metrics elaborated on 3.4, the following null (H_0) and the alternative (H_A) hypothesis is defined:

Null Hypothesis: Using explanation techniques like SHAP, LIME, ANCHORS, and DICE will not determine the best interpretation framework for explaining local credit card fraud classification results.

Alternate Hypothesis: IF a Neural Network algorithm is trained on a credit card transaction dataset for ML fraud detection, and SHAP, LIME, ANCHORS, and DICE interpretability frameworks are applied to individual model results **THEN** a significance test applied to the scores of each interpretability framework, against a predefined set of quality metrics, to rank each explainer technique and statistically determine which is best for explaining local credit card fraud classification results.

To test the hypothesis mentioned above, we conducted a Friedman test to compare four different explainability methods against various evaluation metrics.

Normalization of data is not a prerequisite for the Friedman test, as this non-parametric method is designed to handle data that may not adhere to a normal distribution by comparing ranks rather than actual values. However, it is helpful to perform this transformation to improve the presentation of a graphical analysis. Figure 3 illustrates the model distributions of explainability values (SHAP, LIME, ANCHORS and DiCE) resulting from the test set used to derive the model explainability. In the Box Plot; the x-axis outlines the four XAI methods (SHAP, LIME, ANCHORS, and DiCE), y-axis the score and the box shows the interquartile range (IQR), indicating the middle 50% of scores for each method. We used the same ANN model as a baseline for SHAP, LIME, ANCHORS, and

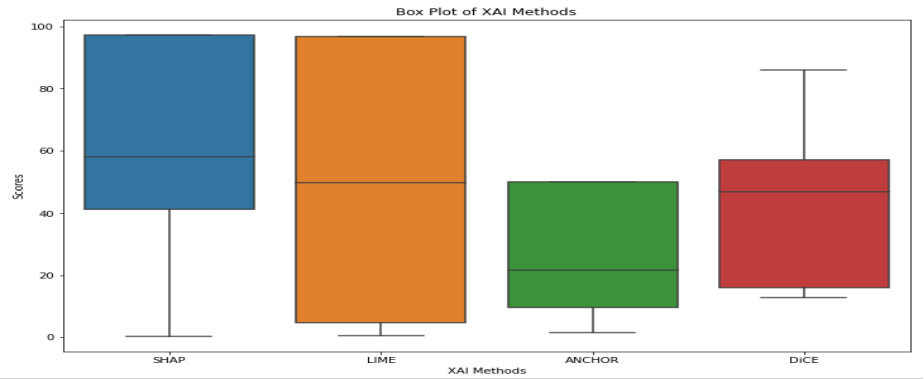


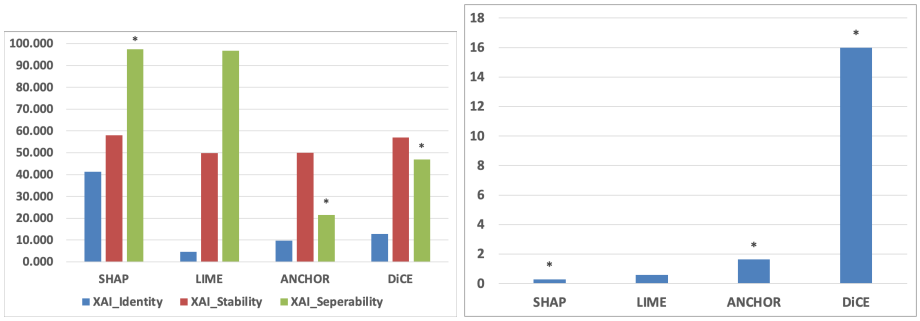
Fig. 3: Box Plot Analysis of Distributions accross XAI methods.

DiCE methods. This allowed for statistical comparisons between the different explainability methods and multiple datasets generated from the same baseline test set. The advantage of using the Friedman test is that it can handle these comparisons efficiently. Table 3 outlines that statistical significance test across different evaluation metrics against the adopted XAI methods (SHAP, LIME, ANCHORS and DiCE). Based on the statistical significance analysis with the Friedman test,

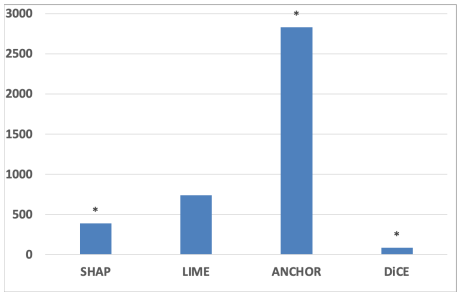
Table 3: Friedman Test of Significance between XAI methods and Explainability metrics (The † indicates the significant result).

	SHAP	**LIME**	**ANCHORS**	**DiCE**
XAI Identity	47.76 (0.0001)†			
XAI Stability	1.56 (0.6849)			
XAI Seperability	54.015 (0.0001)†			
XAI Similairity	60 (0.0001)†			
Comp. Efficiency	60 (0.0001)†			
	test statistic (p-value)			

we can see that except for the Stability metrics, all others are statistically significant against the adopted significance level of $\alpha < 0.005$. By further comparing the XAI methods (SHAP, LIME, ANCHORS and DiCE) with statistically significant metrics (Identity, Separability, Similarity and Computing Efficiency) we can see the best XAI methods. The results indicated that, in terms of Iden-



(a) Identity, Stability & Separability (b) Similarity



(c) Computing Efficiency

Fig. 4: Comparison of XAI methods against XAI evaluation metrics (The $*$ indicate significant XAI method results)

.

tity, SHAP provided the highest identity $(41, 30)$, meaning that it reflected the most identically by capturing the important factors contributing to the model's decision-making process, followed by DiCE $(12, 84)$ and ANCHORS$(9, 86)$. Similarly, in Separability metrics, SHAP $(97, 38)$ still showed the greatest capability to distinguish between different classes or categories in the data, followed by DiCE $(47, 00)$ and ANCHORS $(21, 56)$. Very tight results are witnessed in the Stability metrics where the consistency of explanations produced by an XAI method when perturbations are introduced to the input data provided to the model yielded approximately similar performance between $58, 00$ and $56, 00$ for SHAP and DiCE and $49, 00$ for ANCHORS respectively. Considering the Similarity metrics, we witness that explanations generated by the SHAP method resemble the most to the decision-making process of the ANN model (0.28), fol-

lowed by ANCHOR (1, 62) and DiCE (15, 99). In terms of computing efficiency DiCE demonstrated the better performance than ANCHOR and SHAP.

5 Conclusion

This paper examines the challenge of proposing a quantifiable framework to assess Artificial Intelligence methods (XAI) that explain why a given credit card transaction is labelled as fraudulent. The work provided an objective analysis as to whether a given state-of-the-art ML explainer could be shown to provide the best explanations when compared against other explainer techniques in this particular financial crime domain. Our experiments showed that, after training a Neural Network model on a credit card fraud dataset, it was possible to distinguish between the merits of the SHAP, ANCHORS, and DiCE interpretability methods used in the experiments in this research. The results for LIME were not statistically significant and thus inconclusive.

The key observations from metric scores across the XAI methods can be summarized around the following points:

- The Stability score is the only metric with consistent values across the XAI Methods. On average, 50% of the explanation clusters match the grouping of fraud and non-fraud instances. Variations are more clear-cut across the other metrics.
- Identity scores poorly for all XAI methods, except SHAP. This was arguably the most straightforward metric: similar instances should have similar explanations. However, the SHAP technique provides a score for all instance values, while the other methods generate explainers that only cover some instance attributes. This does not invalidate the use of this metric, but it is a key consideration when assessing this research.
- ANCHORS is one of the methods that produce relatively sparse output (only one feature might be classified as an 'anchor' in the classification result). Thus, the explainers score relatively poorly in distinguishing themselves from each other (Separability).
- The magnitude of counterfactual explanations for different instances can vary significantly. Hence, the DiCE method produces outputs where the Euclidean distance between separate explanations can be significant. This characteristic explains the higher DiCE score for Similarity.
- The inclusion of the Computation Efficiency metric may initially appear discordant within the scope of this study, as it pertains more to overall system performance rather than directly assessing the quality of the explanatory output. However, the speed (or lack thereof) with which an explainer can process this credit card dataset is deemed a critical measure for comparison. DiCE and SHAP generation time is considerably quicker than ANCHORS, for example. Even with parallel/scaleable batch processing options, the user of ANCHORS for explanations in a commercial high-volume fraud detection environment may not be viable.

Bibliography

[1] Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6(52):138–160.

[2] Anowar, F. and Sadaoui, S. (2020). Incremental neural-network learning for big fraud data. *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 1(1):1–4.

[3] Batageri, A. and Kumar, S. (2021). Credit card fraud detection using artificial neural network. *Global Transitions Proceedings*, 2(1):35–41.

[4] Carvalho, D. V., Pereira, E. M., and Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8).

[5] 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.

[6] Darias, J. M., Caro-Martínez, M., Díaz-Agudo, B., and Recio-Garcia, J. A. (2022). Using case-based reasoning for capturing expert knowledge on explanation methods. *Case-Based Reasoning Research and Development*, 13405:3–17.

[7] ElShawi, R., Sherif, Y., Al-Mallah, M., and Sakr, S. (2020). Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Computational Intelligence*, 37(4):1633–1650.

[8] Hailemariam, Y., Yazdinejad, A., Parizi, R., Srivastava, G., and Dehghantanha, A. (2020). An empirical evaluation of ai deep explainable tools. *2020 IEEE Globecom Workshops (GC Wkshps*, pages 1–6.

[9] Hanafy, M. and Ming, R. (2022). Classification of the insureds using integrated machine learning algorithms: A comparative study. *Applied Artificial Intelligence*, 36.

[10] Honegger, M. (2018). Shedding light on black box machine learning algorithms: Development of an axiomatic framework to assess the quality of methods that explain individual predictions.

[11] Ignatiev, A. (2020). Towards trustable explainable ai. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, page 5154–5158.

[12] Jacob, V., Song, F., Stiegler, A., Rad, B., Diao, Y., and Tatbul, N. (2021). Exathlon: A benchmark for explainable anomaly detection over time series. *Proceedings of the VLDB Endowment*, 14(11):2613–2626.

[13] Jesus, S., Belém, C., Balayan, V., Bento, J., Saleiro, P., Bizarro, P., and Gama, J. (2021). How can i choose an explainer? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

[14] Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., and Wortman-Vaughan, J. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, page 1–14.

[15] Lundberg, S. M. and Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*, volume 30. NeurIPS Proceedings.

[16] Moreira, C., Chou, Y., Velmurugan, M., Ouyang, C., Sindhgatta, R., and Bruza, P. (2020). Linda-bn: An interpretable probabilistic approach for demystifying black-box predictive models. *Decision Support Systems*, 150.

[17] Mothilal, R. K., S. A. and Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.*

[18] Nascita, A., Montieri, A., G.Aceto, Ciuonzo, D., Persico, V., and Pescape, A. (2021). Unveiling mimetic: Interpreting deep learning traffic classifiers via xai techniques. *2021 IEEE International Conference on Cyber Security and Resilience (CSR)*, page 455–460.

[19] Nesvijevskaia, A., Ouillade, S., Guilmin, P., and Zucker, J.-D. (2021). The accuracy versus interpretability trade-off in fraud detection model. *Data & Policy*, 3:e12.

[20] Nguyen, M., Bouaziz, A., Valdes, V., Rosa-Cavalli, A., Mallouli, W., and MontesDeOca, E. (2023). A deep learning anomaly detection framework with explainability and robustness. *Proceedings of the 18th International Conference on Availability, Reliability and Security.*

[21] Ras, G., Xie, N., Gerven, M. V., and Doran, D. (2022). Explainable deep learning: A field guide for the uninitiated. *Journal of Artificial Intelligence Research*, 73:329–397.

[22] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "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*, page 1135–1144.

[23] Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

[24] Schlegel, U., Arnout, H., El-Assady, M., Oelke, D., and Keim, D. (2019). Towards a rigorous evaluation of xai methods on time series. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW).*

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

[26] Sharma, P. and Priyanka, S. (2020). Credit card fraud detection using deep learning based on neural network and auto encoder. *International Journal of Engineering and Advanced Technology*, 9(5):1140–1143.

[27] Sinanc, D., Demirezen, U., and Sağıroğlu, (2021). Explainable credit card fraud detection with image conversion. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, 10(1):63–76.

[28] Sullivan, R. and Longo, L. (2023). Explaining deep q-learning experience replay with shapley additive explanations. *Machine Learning and Knowledge Extraction*, 5(4):1433–1455.

[29] T.Y.Wu and 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).*

[30] Vilone, G. and Longo, L. (2021a). Classification of explainable artificial intelligence methods through their output formats. *Machine Learning and Knowledge Extraction*, 3(3):615–661.

[31] Vilone, G. and Longo, L. (2021b). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106.

[32] Vouros, G. (2022). Explainable deep reinforcement learning: State of the art and challenges. *ACM Computing Surveys*, 55(5):1–39.

[33] Y, S. and Challa, M. (2023). A comparative analysis of explainable ai techniques for enhanced model interpretability. *2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN)*, pages 229–234.