A comparative analysis of SHAP, LIME, ANCHORS, and DICE for interpreting a dense neural network in Credit Card Fraud Detection

Bujar Raufi$^{1[1111-2222-3333-4444]}$, Ciaran Finnegan$^{1[0009-0008-9620-5460]}$, and Luca Longo$^{1[000-0002-2718-5426]}$

Artificial Intelligence and Cognitive Load Research Lab,
School of Computer Science,
Technological University Dublin,
Dublin, Republic of Ireland
bujar.raufi,luca.longo@tudublin.ie

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 early. 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 · methods comparison · SHapley Additive exPlanations (SHAP) · Local Interpretable Model-agnostic Explanation · ANCHORS · Diverse Counterfactual Explanations

1 Introduction

Credit card fraud costs the financial services industry billions of Euros in losses each year [30] [3]. The need for ever more sophisticated Machine Learning approaches to tackle this problem has been well established in the academic community [11], [38]. Research outlined in [2] and [8] is an example of work in this field to improve fraud detection rates through 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 [41]. This problem is also identified in a recent manifesto on eXplaibable Artificial Intelligence [24], whereby AI-based solutions are used at scale to detect fraud, diagnose investment portfolios for different optimization purposes and predict credit risk [26]. The need for a more trustworthy AI, as well as the transparency of automated decision-making, is gaining momentum as it becomes part of European legal demands [20], [10]. The notion of 'trust' is not new, as many computational trust and reputation models exist in the literature [13]. However, trust is an ill-defined concept and assessing it objectively in current AI-based technologies is not trivial [16]. Fortunately, eXplainable Artificial Intelligence, an emerging interdisciplinary field of research, is focused on developing methods for understanding opaque AI-based models developed within the larger field of Artificial Intelligence for diverse real-world applications. In turn, the development of such methods also helps develop AI-based technologies that are assumed to be more trustworthy, with practical and ethical benefits across various domains. Producing an objective assessment of state-of-the-art ML explainers applied to credit card fraud detection is essential [33]. This current research intends to compare various XAI techniques and look for insights into each one's relative strengths in the context of fraud detection. The experiment focuses on applying XAI methods in a commercial dataset containing credit card transactions, labelled into two classes indicating regular and fraudulent activities.

In detail, this research focuses on the explainability of machine learning-driven software used within 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 research utilises a series of metrics to rate the performance of four state-of-the-art XAI methods: **SH**apley **A**dditive ex**P**lanations (SHAP), **L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations, ANCHORS, and **Di**verse **C**ounterfactual **E**xplanations (DiCE). These methods are explicitly applied to the classification of individual credit card transactions. 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 research addresses the following research question: *'To what extent can some of the explanation techniques provide better explanations on the classification of credit card fraud transactions by a 'black box' Artificial Neural Network (ANN) Machine Learning Model?'*

The remainder of this research manuscript 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 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 [38] [7]. 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 [5]. For example, better ways are needed to assess model outputs and account for their computational efficiency at runtime [4]. 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 inferential process, represents a challenge [35] [15]. 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 [14], [19], [39].

Despite efforts to develop universal frameworks for interpreting Machine Learning (ML) models and their predictions [25], 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 and explainability [1, 43]. 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 [42]. 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 [21]. It is also desirable for humans taking part in XAI-based explanations to have some domain knowledge. However, fraud detection experiments for explainability showed that this can still be subject to user bias [22]. 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 [23]. In line with this, the comparative experiment proposed in this research study will focus on generating model explanations without direct human involvement to avoid bias and provide more programmatic experiments with quantifiable metrics[12].

Much of the research on automated model explanation can be seen in anomaly detection frameworks for explainability [31] and XAI applied to time-series, both explained with the SHAP and LIME explainers [37]. Furthermore, a framework using local explanations generated with SHAP, LIME, and Counterfactual techniques can be seen in [27] and [18]. A significant body of research on XAI approaches is witnessed in using Deep Learning (DL) techniques such as Deep SHAP [29], [40], LIME for Deep Neural Network (DNN) experiments [34], and the use of counterfactuals in CNN model explainers [45].

The **SH**apley **A**dditive ex**P**lanations (SHAP) framework is a method that helps in interpreting predictions of data-driven models [25]. It is derived from cooperative game theory, based on Shapley values and is widely used to understand how input features relate to a model's prediction. Shapley values are represented as:

$$\phi_i^s = \frac{1}{\Pi(\mathcal{F})} \sum_{\pi \in \Pi(\mathcal{F})} [\upsilon(\mathcal{P}_i^\pi \cup \{i\}) - \upsilon(\mathcal{P}_i^\pi)] \tag{1}$$

where, the expression inside the sum represents the i^{th} features marginal contribution within permutation π. According to the equation, the Shapley value for a feature is the average marginal contribution of that feature calculated across all possible permutations of the feature set. 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 the inferential process learnt by black box models [35]. The explanation of an instance d, which is subject to interpretation, can be represented as

$$e(d) = argmin_{g \in G} \xi(f, g, \pi_d) \cdot \Omega(g) \tag{2}$$

where d might be a new instance, given that it can be adequately represented within the framework of the training data utilized by the black box model. The definition of d originates from the maximization of a fidelity term affiliated to loss, denoted as $\xi(f, g, \pi_d)$, along with a complexity term denoted as $\Omega(g)$. Here, f denotes the black box model under scrutiny, while g denotes the explanatory model. G represents the comprehensive hypothesis space of a specific interpretable model. Upon minimization, the explanation must handle two crucial trade-off terms: $\xi(f, g, \pi_d)$ tends to achieve an optimal alignment of g with the model f, whilst a lower loss tends to maintain higher fidelity within the local context. The attainment of optimal alignment depends on proximity measure denoted as $\pi_d(z)$ within the vicinity of d. The main idea behind LIME is to identify the input features that have the highest impact on predicting a specific target class of interest within a trained 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. Another class of explainers is based on rules, considered more interpretable than numerical outputs [44].

ANCHORS is a model-agnostic explanation approach that provides interpretable rules for the model's predictions. It offers local and global explanations, introduced by Ribeiro [36]. Given the black-box model f, an instance d, a distribution $\mathcal{D}$ together with a desired precision τ, we can define a precision of an anchor $\aleph$ as a set of feature predicates d exhorting $prec(\aleph) \geq \tau$ given as:

$$prec(\aleph) = \mathbb{E}_{\mathcal{D}(z|\aleph)}[\mathbb{1}_{f(d)=f(z)}] \tag{3}$$

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 [28]. DICE offers actionable insights by identifying the minimum changes required to modify a model's prediction for a specific instance. One of the methods for minimising the changes or losses was introduced in [46] and is given as:

$$\mathcal{L}(x, x', y, y', \lambda) = \lambda \cdot (\hat{f}(x') - y')^2 + d(x, x') \tag{4}$$

The first term in the equation represents the quadratic distance between the model prediction for the counterfactual x' and the desired outcome y', which must be defined beforehand. The second term is the distance d between the instance x to be explained and the counterfactual x'. The loss function measures the difference between the forecasted outcome of the counterfactual and the predetermined outcome and the difference between the counterfactual and the instance of interest. This distance metric, denoted as d, is quantified by a Manhattan distance weighted by the inverse median absolute deviation (MAD) of individual features, given as:

$$d(x, x') = \sum_{j=1}^{p} \frac{|x_j - x'_j|}{MAD_j} \tag{5}$$

The method helps to address the common challenge of opaque decision-making in complex models. Still, it 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 [47, 17]. 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, this research identifies a significant void in contemporary research regarding establishing a comprehensive framework for explaining credit card fraud classifications made by 'black-box' models [43]. 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 and methodology

The research aim is to compare four interpretability frameworks (LIME, SHAP, Anchors, and DiCE), using predefined, custom-built comparison metrics, against the output of a Neural Network (NNet) credit card fraud detection model. The goal is to establish if one XAI method consistently rank higher than the other methods. The design pipeline includes a phase for preparing and pre-processing the dataset, hyperparameters tuning of a Dense Network, model building via training, model evaluation, and model explanation, as depicted in figure 1.

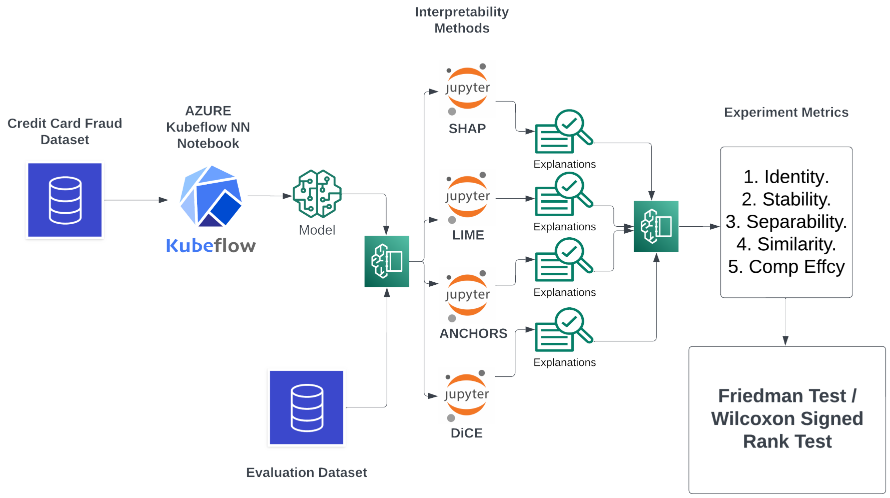


Fig. 1: Overview of the design of a comparative experiment of LIME, SHAP, Anchors, and DiCE

3.1 Dataset and Pre-processing

The dataset utilized in this study was provided by a private company and is associated with a product development cycle 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, each 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 dataset. Approximately 15 % of the records in this dataset are fraudulent, which is more balanced than typical credit card fraud datasets [32]. However, a down-sampling of the non-fraudulent records is applied to minimise biases, create an even classification split, and train a fairer Neural Network for predictive purposes. A portion of non-fraudulent records is removed to simplify the process and avoid adding new synthetic data. This is implemented using a randomly down-sampling Python function. As a result, the remaining dataset has only $7,000$ rows and a 50%/50% breakdown of fraudulent and non-fraudulent records. The initial dataset comprises 380 features, and AzureML's permutation feature importance is used to reduce the attribute list to key columns. Permutation Feature Importance assesses the significance of features by quantifying the model's responsiveness to random permutations of their values. This methodology assumes that perturbing the values of pivotal features induces a clearer decline in model performance contrary to their less influential counterparts [9]. The final dataset consists of $7,000$ rows and 64 features. The complete list of final features, their type and value range is provided in table 5 of the appendix.

3.2 Model building, hyperparameter tuning and training

A feed-forward artificial neural network (ANN) with dense layers is employed to classify fraudulent transactions. Hyperparameter tuning is used to design an optimal architecture for training. It uses a random search method across six hyperparameters spread across the ANN's input and hidden layers. Table 1 outlines such hyperparameters, the search space and the selected value for the architecture. Model training is done up to 100 epochs using the Adam optimization algorithm with the binary cross-entropy loss function. A typical 80/20 split is employed for train and testing.

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

3.3 Model evaluation

In various experiments present in the literature, conducted to detect fraud using Artificial Neural Networks (ANN), an accuracy score of $>= 0.85$ and an F1 score of $>= 0.85$ have been identified as ideal targets [39],[6]. 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, the model's overall accuracy score measures 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 a 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 to 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

A single predictive model for credit card fraud is built and saved to assess the explanations from each XAI selected method (SHAP, LIME, ANCHORS and DICE). However, generating a number of the XAI explanations is a processor-intensive endeavour, and even attempting to generate both explanations and metrics on the single $1.4K$ test data block was found to be impractical and prone to system timeouts. Thus, the test data is subsequently split 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 doing.
2. *Stability* - Instances belonging to the same class have comparable explanations. K-means clustering is applied to explain each instance in test data. It measures the number of explanations in both clusters (fraud/non-fraud) that match the predicted class. A higher stability score implies that the explanations remain consistent across different scenarios, which is crucial for building trust in the developed 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 inferential 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). It clusters test data instances into Fraud/non-Fraud clusters. It also normalises explanations and calculates Euclidean distances between instances in both clusters. Smaller mean pairwise distance equalizes to a better explainability metric.

5. *Computational time efficiency* - Is the average time taken, in seconds, by the interpretability framework to output a set of explanations. The Kubeflow environment is expected to reduce unexpected factors affecting computational efficiency.

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.

3.5 Synthesis of experimental design

In synthesis, this research is built on the following eight steps to compile a table of metrics for statistically evaluating the selected XAI methods:

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. An XAI method must demonstrate that it can generate explanations on the credit card dataset used in this research.
3. Refine the credit card fraud model-building process if that improves the quality of the explanations without compromising on model performance. For example, the DiCE XAI method will require a feature set with sufficient continuous attributes to allow a counterfactual search space to generate explanations.
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. Working with blocks of test data is necessary because some of the XAI methods are computationally very heavy, and processing the entire test dataset all at once was impractical.
5. Submit each XAI output data block to a separate Python function to generate a value from each experiment metric (Identity, Stability, Separability, Similarity, and Computation Efficiency).
6. Take the metric scores of each block of data for each XAI technique and use these values as input for a statistical significance comparison.
7. Conduct a comparative statistical analysis of the XAI metric score for each XAI method to determine if any significant difference in performance exists.

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

> **IF** a dense Neural Network, optimally tuned, is trained on a balanced credit card transaction dataset for fraud detection. SHAP, LIME, ANCHORS, and DICE XAI methods are applied to the individual model results for the Test Set **THEN** there exists at least one of such XAI methods that rank significantly higher across a set of explanation metrics (identity, stability, separability, computational efficiency).

To test the hypothesis, a Friedman test compares the four explainability methods (SHAP, LIME, ANCHORS, and DICE) against five evaluation metrics (identity, stability, separability, computational efficiency). 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.

In the tabular representation of this Friedman Test, as seen in Table 4, the 'Statistic' column refers to the Friedman statistic, quantifying the differences among group ranks across multiple related samples or matched groups. The 'p-value' column represents the probability that the observed differences among the ranks could have occurred by chance under the null hypothesis that all groups have the same distribution.

4 Results and discussion

This section presents the main findings of the comparative experiments designed in section 3

4.1 Data Preprocessing findings

The credit card transaction dataset consisted of two labels in the target class: non-fraud (0) and fraud (1). Table 2 illustrates the data preparation and pre-processing activities done against the dataset regarding instances and features, removal of redundant features, outliers and the number of positive and negative train and test examples.

Table 2: Pre-processing activities executed on the dataset

Activity	**Value**
Initial instances	25128
Initial Features	380
Rows Removed (Deleted rows with high volumes of missing data elements)	297
Outliers removed (Deleted exceptionally high trxn amounts)	1240
Number of Features (After highly correlated/redundant features removal)	64
Number of continuous features	59
Number of categorical features	5
Number of train instances	5256
Number of positive train instances	2635
Number of negative train instances	2621
Number of test instances	1314
Number of positive test instances	650
Number of negative test instances	664

4.2 Evaluating the Predictive Fraud Model

This predictive credit card fraud model trained in this research experiment was required to demonstrate strong performance across the selected evaluation metrics (Accuracy, ROC-AUC, Precision, Recall, F1-score). Only if this is confirmed would a meaningful evaluation of the selected XAI methods make sense. After a series of iterations, including refinements required to improve the quality of the XAI experiment results, a credit card predictive model was created with the evaluation metrics generated on the Test Set as shown in the table figure 3 below.

Table 3: Model Performance Metrics

Accuracy	0.86
ROC AUC Score	0.93
Precision (Class 0)	0.92
Recall (Class 0)	0.79
F1-Score (Class 0)	0.85
Precision (Class 1)	0.81
Recall (Class 1)	0.93
F1-Score (Class 1)	0.87

4.3 Model Explanations and Comparisons

Figure 2 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), the 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,

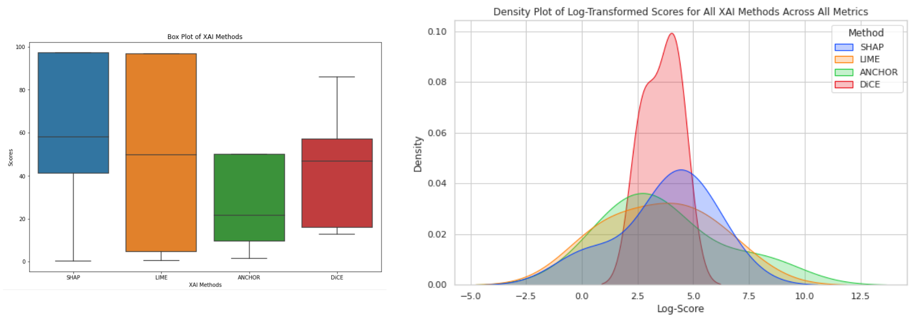


Fig. 2: Box Plot / Density Analysis of Distributions across XAI methods.

ANCHORS, and 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 4 outlines that statistical significance test across different evaluation metrics against the adopted XAI methods (SHAP, LIME, ANCHORS and DiCE).

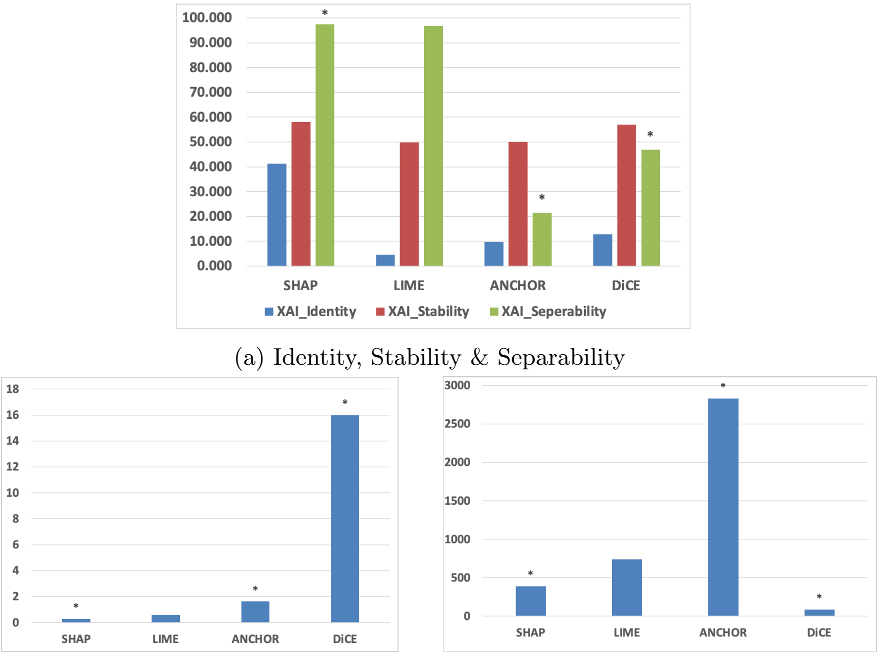
Table 4: Friedman Test of Significance between XAI methods and Explainability metrics

Metric	Statistic	P-Value	Significant Difference
Identity	47.76	2.40e-10	Yes
Stability	1.56	0.668493	No
Separability	55.4	5.64e-12	Yes
Similarity	60.00	5.88e-13	Yes
Computational Efficiency	60.00	5.88e-13	Yes

Based on the statistical significance analysis with the Friedman test, 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 Identity, 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), followed by ANCHOR $(1, 62)$ and DiCE $(15, 99)$. Regarding computing efficiency, DiCE demonstrated better performance than ANCHOR and SHAP.

In synthesis, 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 selected 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.



(a) Identity, Stability & Separability

(b) Similarity

(c) Computing Efficiency

Fig. 3: Comparison of XAI methods against XAI evaluation metrics (The $*$ indicate significant XAI method results)

.

- 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 method that produces relatively sparse output. The classification results might classify only one feature as an 'anchor'. Thus, according to the Separability metric, such an explainer scored relatively poorly, distinguishing itself from the others.
- 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 an important measure for com-

parison, as it is connected to actionable XAI in real-world fields. DiCE and SHAP generation time is considerably quicker than ANCHORS. Even with parallel/scaleable batch processing options, using ANCHORS for explanations in a commercial high-volume fraud detection environment may not be viable.

5 Conclusion

This research endeavoured to establish a quantitative framework for assessing various Explainable Artificial Intelligence (XAI) methods in the context of credit card fraud detection. The study utilized a neural network model trained on a dataset of credit card transactions, evaluating the efficacy of XAI techniques like SHAP, LIME, ANCHORS, and DiCE against multiple custom metrics. While the results did not conclusively favour any single technique, they highlighted the nuanced performance of these methods across different metrics, demonstrating the complexity of applying XAI in practical scenarios. The XAI techniques scored differently on four of the five custom metrics. An observer could declare one technique to be the *best*, but only if a weighting was applied to one or more of the metric scores.

The findings revealed varied performances across the XAI methods, with SHAP generally excelling in identity and similarity metrics. At the same time, ANCHORS struggled with computational efficiency, significantly impacting its practical applicability in high-volume environments. The experiments underscored the importance of choosing appropriate metrics for evaluating XAI techniques and the challenges of applying these methods across different domains and data characteristics.

This work contributes to the broader understanding of XAI in financial services by providing a structured evaluation of multiple explanation techniques. It underlines the potential of these methods to enhance transparency in AI-driven decisions, which is crucial for sectors like financial crime detection. The study's methodology and findings offer a foundation for future research to build upon, particularly in exploring the adaptability of XAI methods across various applications.

Future work should expand the range of XAI techniques and metrics evaluated, exploring their applicability in broader datasets and different contexts. Integrating Generative AI to provide narrative explanations alongside traditional XAI outputs could further enhance the interpretability of AI decisions, adapting to the evolving demands of AI transparency in industry practices. This approach would refine the utility of XAI and push the boundaries of what these technologies can achieve in practical, high-stakes environments.

Disclosure of Interests. The authors have no competing interests to declare relevant to this article's content.

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (xai). IEEE Access **6**(52), 138–160 (2018). https://doi.org/10.1109/access.2018.2870052
2. Ajitha, E., Sneha, S., Makesh, S., Jaspin, K.: A comparative analysis of credit card fraud detection with machine learning algorithms and convolutional neural network. In: 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI). pp. 1–8 (2023). https://doi.org/10.1109/ACCAI58221.2023.10200905
3. Alam, M.N., Podder, P., Bharati, S., Mondal, M.R.: Effective machine learning approaches for credit card fraud detection. advances in intelligent systems and computing. Innovations in Bio-Inspired Computing and Applications. IBICA 2020 (2021). https://doi.org/doi:10.1007/978-3-030-73603-3_14
4. Alarfaj, F.K., Iqra Malik, I., Ullah Khan, H., Muhammad Ramzan, N.A., Ahmed, M.: Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms. IEEE Access (2022). https://doi.org/10.1109/ACCESS.2022.3166891
5. Alvarez-Melis, D., Jaakkola, T.: On the robustness of interpretability methods. 2018 ICML Workshop on Human Interpretability in Machine Learning (2018). https://doi.org/10.48550/arXiv.1806.08049
6. Anowar, F., Sadaoui, S.: Incremental neural-network learning for big fraud data. 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC) **1**(1), 1–4 (2020). https://doi.org/10.1109/smc42975.2020.9283136
7. Aurna, N.F., Hossain, M.D., Taenaka, Y., Kadobayashi, Y.: Federated learning-based credit card fraud detection: Performance analysis with sampling methods and deep learning algorithms. In: 2023 IEEE International Conference on Cyber Security and Resilience (CSR). pp. 180–186 (2023). https://doi.org/10.1109/CSR57506.2023.10224978
8. Batageri, A., Kumar, S.: Credit card fraud detection using artificial neural network. Global Transitions Proceedings **2**(1), 35–41 (2021). https://doi.org/10.1016/j.gltp.2021.01.006
9. Breiman, L.: Random forests. Machine learning **45**, 5–32 (2001)
10. Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: A survey on methods and metrics. Electronics **8**(8) (Jul 2019). https://doi.org/10.3390/electronics8080832
11. Dal Pozzolo, A., et al.: Learned lessons in credit card fraud detection from a practitioner perspective. Expert Systems with Applications **41**(10), 4915–4928 (2014). https://doi.org/10.1016/j.eswa.2014.02.026
12. Darias, J.M., Caro-Martínez, M., Díaz-Agudo, B., Recio-Garcia, J.A.: Using case-based reasoning for capturing expert knowledge on explanation methods. Case-Based Reasoning Research and Development **13405**, 3–17 (Aug 2022). https://doi.org/10.1007/978-3-031-14923-8_1
13. Dondio, P., Longo, L.: Trust-based techniques for collective intelligence in social search systems. In: Next generation data technologies for collective computational intelligence, pp. 113–135. Springer (2011)
14. ElShawi, R., Sherif, Y., Al-Mallah, M., Sakr, S.: Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. Computational Intelligence **37**(4), 1633–1650 (Aug 2020). https://doi.org/10.1111/coin.12410

15. Evans, B.P., Xue, B., Zhang, M.: What's inside the black-box? Proceedings of the Genetic and Evolutionary Computation Conference (Jul 2019). https://doi.org/10.1145/3321707.3321726
16. Fritz Morgenthal, S., Hein, B., Papenbrock, J.: Financial risk management and explainable, trustworthy, responsible ai. Frontiers in Artificial Intelligence (Feb 2022). https://doi.org/https://doi.org/10.3389/frai.2022.779799
17. Hailemariam, Y., Yazdinejad, A., Parizi, R., Srivastava, G., Dehghantanha, A.: An empirical evaluation of ai deep explainable tools. 2020 IEEE Globecom Workshops (GC Wkshps pp. 1–6 (2020). https://doi.org/10.1109/GCWkshps50303.2020.9367541
18. Hanafy, M., Ming, R.: Classification of the insureds using integrated machine learning algorithms: A comparative study. Applied Artificial Intelligence **36** (2022). https://doi.org/10.1080/08839514.2021.2020489
19. Honegger, M.: Shedding light on black box machine learning algorithms: Development of an axiomatic framework to assess the quality of methods that explain individual predictions (Aug 2018), https://arxiv.org/abs/1808.05054v1
20. Ignatiev, A.: Towards trustable explainable ai. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence p. 5154–5158 (Jul 2020). https://doi.org/10.24963/ijcai.2020/726
21. Jacob, V., Song, F., Stiegler, A., Rad, B., Diao, Y., Tatbul, N.: Exathlon: A benchmark for explainable anomaly detection over time series. Proceedings of the VLDB Endowment **14**(11), 2613–2626 (2021). https://doi.org/10.14778/3476249.3476307
22. Jesus, S., Belém, C., Balayan, V., Bento, J., Saleiro, P., Bizarro, P., Gama, J.: How can i choose an explainer? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (Mar 2021). https://doi.org/10.1145/3442188.3445941
23. Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., Wortman-Vaughan, J.: Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems p. 1–14 (Apr 2020). https://doi.org/10.1145/3313831.3376219
24. Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Ser, J.D., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Páez, A., Samek, W., Schneider, J., Speith, T., Stumpf, S.: Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. Information Fusion **106**, 102301 (2024). https://doi.org/https://doi.org/10.1016/j.inffus.2024.102301, https://www.sciencedirect.com/science/article/pii/S1566253524000794
25. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). vol. 30. NeurIPS Proceedings (2017)
26. Misheva, B., Osterrieder, J., Hirsa, A., Kulkami, O., Lin, S.F.: Explainable ai in credit risk management (2021). https://doi.org/https://doi.org/10.48550/arXiv.2103.00949
27. Moreira, C., Chou, Y., Velmurugan, M., Ouyang, C., Sindhgatta, R., Bruza, P.: Linda-bn: An interpretable probabilistic approach for demystifying black-box predictive models. Decision Support Systems **150** (2020). https://doi.org/10.1016/j.dss.2021.113561
28. Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. Proceedings of the

2020 Conference on Fairness, Accountability, and Transparency (2020). https://doi.org/10.48550/arXiv.1806.08049

29. Nascita, A., Montieri, A., G.Aceto, Ciuonzo, D., Persico, V., Pescape, A.: Unveiling mimetic: Interpreting deep learning traffic classifiers via xai techniques. 2021 IEEE International Conference on Cyber Security and Resilience (CSR) p. 455–460 (2021). https://doi.org/10.1109/csr51186.2021.9527948
30. Nesvijevskaia, A., Ouillade, S., Guilmin, P., Zucker, J.D.: The accuracy versus interpretability trade-off in fraud detection model. Data & Policy **3**, e12 (2021)
31. Nguyen, M., Bouaziz, A., Valdes, V., Rosa-Cavalli, A., Mallouli, W., MontesDeOca, E.: A deep learning anomaly detection framework with explainability and robustness. Proceedings of the 18th International Conference on Availability, Reliability and Security. (2023). https://doi.org/10.1145/3600160.3605052
32. Priscilla, C., Prabha, D.: Influence of optimizing xgboost to handle class imbalance in credit card fraud detection. 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT) p. 1309–1315 (2020). https://doi.org/10.1109/icssit48917.2020.9214206
33. Psychoula, I., Gutmann, A., Mainali, P., Lee, S.H., Dunphy, P., Petitcolas, F.: Explainable machine learning for fraud detection. Computer **54**(10), 49–59 (2021). https://doi.org/10.1109/mc.2021.3081249
34. Ras, G., Xie, N., Gerven, M.V., Doran, D.: Explainable deep learning: A field guide for the uninitiated. Journal of Artificial Intelligence Research **73**, 329–397 (2022). https://doi.org/10.1613/jair.1.13200
35. Ribeiro, M.T., Singh, S., Guestrin, C.: "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 p. 1135–1144 (Aug 2016). https://doi.org/10.1145/2939672.2939778
36. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations. Proceedings of the AAAI Conference on Artificial Intelligence **32**(1) (Feb 2018). https://doi.org/10.1609/aaai.v32i1.11491
37. Schlegel, U., Arnout, H., El-Assady, M., Oelke, D., Keim, D.: Towards a rigorous evaluation of xai methods on time series. 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). (2019). https://doi.org/10.1109/iccvw.2019.00516
38. Sharma, P., Priyanka, S.: 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 (Jun 2020). https://doi.org/10.35940/ijeat.e9934.069520
39. Sinanc, D., Demirezen, U., Sagıroglu, S.: Explainable credit card fraud detection with image conversion. ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal **10**(1), 63–76 (2021). https://doi.org/10.14201/adcaij20211016376
40. Sullivan, R., Longo, L.: Explaining deep q-learning experience replay with shapley additive explanations. Machine Learning and Knowledge Extraction **5**(4), 1433–1455 (2023). https://doi.org/10.48550/arXiv.1806.08049
41. T.Y.Wu, Y.T.Wang: Locally interpretable one-class anomaly detection for credit card fraud detection. 2021 International Conference on Technologies and Applications of Artificial Intelligence (TAAI) (2021). https://doi.org/10.1109/taai54685.2021.00014
42. Vilone, G., Longo, L.: Classification of explainable artificial intelligence methods through their output formats. Machine Learning and Knowledge Extraction **3**(3), 615–661 (2021). https://doi.org/10.3390/make3030032

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

44. Vilone, G., Longo, L.: A quantitative evaluation of global, rule-based explanations of post-hoc, model agnostic methods. Frontiers in Artificial Intelligence **4**, 160 (2021). https://doi.org/10.3389/frai.2021.717899, https://www.frontiersin.org/article/10.3389/frai.2021.717899

45. Vouros, G.: Explainable deep reinforcement learning: State of the art and challenges. ACM Computing Surveys **55**(5), 1–39 (2022). https://doi.org/10.1145/3527448

46. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv. JL & Tech. **31**, 841 (2017)

47. Y, S., Challa, M.: A comparative analysis of explainable ai techniques for enhanced model interpretability. 2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN) pp. 229–234 (2023). https://doi.org/10.1109/ICPCSN58827.2023.00043

Appendix

Table 5: Independent variables in the dataset used in the experiment

Feature	**Type**	**Value Range**
OnlinePOSCount.cnt.day.present	continuous	-81 … 29
EMVTransactionsCount.cnt.day.present	continuous	0 … 12
NonEMVTransactionsCount.cnt.day.present	continuous	0 … 81
MerchantCategory	continuous	742 … 9402
POS_Count.cnt.day.present	continuous	0 … 29
PinIndicator	categorical	0 … 1
DomesticAuthCount.cnt.hour1	continuous	0 … 18
DomesticAuthCount.cnt.hour3	continuous	0 … 24
DomesticAuthCount.cnt.hour4	continuous	0 … 28
DomesticAuthCount.cnt.hour10	continuous	0 … 28
DomesticAuthCount.cnt.hour15	continuous	0 … 29
DomesticAuthCounter.cnt.day.present	continuous	0 … 29
DomesticAuthCount.cnt.hour25	continuous	0 … 36
OnlinePOSCountForever.cnt.present	continuous	-385 … 199
POSTerminalAttendedAuthCount.cnt.day.present	continuous	0 … 29
CustomerNotPresentAuthCount.cnt.day.present	continuous	0 … 78
DvcVerificationCap	continuous	0 … 8
ECommerceAuthCount.cnt.day.present	continuous	0 … 67
PosTerminalAttended	categorical	Y, N
TxnChannelCode	categorical	OnL, POS
CustomerPresentIndicator	categorical	Y, N
OnlineNewMerchCtryCntDaily.cnt.day.present	continuous	0 … 9
OnlineNewMerchCtryCntHourly.cnt.hour24	continuous	0 … 9
OnlineNewMerchCtryCntHourly.cnt.hour15	continuous	0 … 9
OnlineNewMerchCtryCntHourly.cnt.hour10	continuous	0 … 9
DvcPosEntryMode	categorical	0 … 9
OnlineNewMerchCtryCntHourly.cnt.hour3	continuous	0 … 9
OnlineNewMerchCtryCntHourly.cnt.hour4	continuous	0 … 9
OnlineNewMerchCtryCntDaily.cnt.day.total	continuous	0 … 29
NotECommerceAuthCount.cnt.day.present	continuous	0 … 425
NewMerchantCountryCount.cnt.hour15	continuous	0 … 9
OnlineNewMerchCtryCntHourly.cnt.hour1	continuous	0 … 425
NewMerchantCountryCount.cnt.hour10	continuous	0 … 9
NewMerchantCountryCount.cnt.hour24	continuous	U … Y
ECommerceFlag	categorical	0 … 9
NewMerchantCountryCount.cnt.hour4	continuous	0 … 9
NewMerchantCountryCount.cnt.hour3	continuous	0 … 903
AuthResponse	categorical	0 … 9
NewMerchantCountryCount.cnt.hour1	continuous	0 … 22885
AmountBase	continuous	1 … 819
CardType	categorical	0 … 20963
POSSum.acc.month.total	continuous	0 … 58468
NotECommerceAuthAmount.acc.day.total	continuous	0 … 63169
NonEMVTransactionsAcc.acc.day.total	continuous	0 … 49670
POSTerminalAttendedAuthAmount.acc.day.total	continuous	0 … 20091
CustomerPresentAuthAmount.acc.day.total	continuous	0 … 11752
EMVTransactionsAcc.acc.day.total	continuous	0 … 63169
CustomerNotPresentAuthAmount.acc.day.total	continuous	0 … 63161
HourlyAuthAmt.acc.hour25	continuous	0 … 63161
NonEMVTransactionsAcc.acc.day.present	continuous	0 … 22885
NotECommerceAuthAmount.acc.day.present	continuous	0 … 63161
CustomerNotPresentAuthAmount.acc.day.present	continuous	0 … 22885
POSTerminalAttendedAuthAmount.acc.day.present	continuous	0 … 12953
CustomerPresentAuthAmount.acc.day.present	continuous	0 … 13068
HighRiskPOSSum.acc.hour.total	continuous	0 … 10957
EMVTransactionsAcc.acc.day.present	continuous	0 … 1