Journal of the Brazilian Computer Society, 2026, 32:1, doi: 10.5753/jbcs.2026.5376

Evaluation of explainable artificial intelligence techniques in the context of credit card fraud detection

Gabriel Mendes de Lima [**Universidade Federal do ABC** | ***mendes.gabriel@ufabc.edu.br***]
Paulo Henrique Pisani [**Universidade Federal do ABC** | ***paulo.pisani@ufabc.edu.br***]

Centro de Matemática, Computação e Cognição (CMCC), Universidade Federal do ABC (UFABC) – Santo André, SP, Brazil.

Received: 28 December 2024 • **Accepted:** 21 November 2025 • **Published:** 25 March 2026

Abstract Artificial intelligence has been employed in several applications in the financial sector. This paper deals with one of these applications: fraud detection in credit card transactions. In this context, a number of machine learning algorithms can be used to obtain models which automate the classification of a transaction as fraudulent or genuine. However, some of these machine learning algorithms are not directly interpretable. The current paper presents an evaluation of explainable artificial intelligence techniques SHAP and LIME applied to models for fraud detection in credit card transactions. Along with the results of the evaluation, the paper discusses the effectiveness and need for explainable artificial intelligence techniques. This paper extends a previous paper by including hyperparameter tuning, new results and an evaluation of the processing time to obtain explanations. The reported results suggest that SHAP obtains better results than LIME, although LIME required less processing time after obtaining the LIME explainer.

Keywords: machine learning, explainable artificial intelligence, credit card fraud detection

1 Introduction

Machine learning algorithms have been applied for financial fraud detection [Aros *et al*., 2024]. It includes the use of machine learning models for detecting frauds in credit card transactions [Makki *et al*., 2019; Khyati Chaudhary, 2012]. However, several machine learning models do not easily provide an explanation for their decisions. When using automated learning in complex scenarios, such as the financial sector, predictive performance is important, as well as other factors. A critical factor for the safe use of machine learning is providing explanations [Doshi-Velez and Kim, 2017]. In this context, interpretability can be defined as the level of understanding which a human being has from the decisions taken by a model [Miller, 2019].

Financial fraud detection is a key application area for machine learning, particularly in view of different types of fraud. These frauds can be internal or external. Examples of internal attacks involve financial statements or money laundering, while external frauds can occur in credit cards and insurance [Aros *et al*., 2024]. During the COVID-19 pandemic, online financial services became more important [Psychoula *et al*., 2021]. In 2019, the amount of money in online financial fraud was 80% greater than the UK's GDP [Gee *et al*., 2019].

Additionally, several decisions in the financial market can be taken by models which were obtained by automated learning. For example, evaluation of loan requests, acceptance of credit card transactions, etc. However, the use of models which take decisions that cannot be explained is not suitable for highly regulated systems like those in the banking sector [Bussmann *et al*., 2021]. The European Union and its member countries enforce the need for providing interpretation that can be understood by a human being [European Union, 2016]. Moreover, the trustworthy of artificial intelligence (AI) involve some ethical questions such as transparency and accountability [Bussmann *et al*., 2021].

Explainable artificial intelligence can provide key contributions in this context. Techniques and models that make the behavior and prediction of machine learning models understandable for humans are part of what is known as explainable artificial intelligence [Molnar, 2022]. Explainable AI techniques can contribute to audit learned models and provide a better understanding of the obtained results [Hanif, 2021]. In the literature, Explainable AI has been applied to Finance as presented in recent reviews [Martins *et al*., 2024; Černevičienė and Kabašinskas, 2024].

In the financial sector, the detection of frauds in credit card transactions is an important application. Fraud in this case refers to the use of credit cards to illegally obtain goods or services [Khyati Chaudhary, 2012]. In this work, the problem is defined in the following way: given a dataset containing past transactions already labeled as fraud or genuine, obtain a model capable of classifying new transactions as fraud or genuine. Some explainable artificial intelligence techniques previously applied in the context of credit card transactions are LIME (*Local Interpretable Model-agnostic Explanations*) and SHAP (*Shapley Additive Explanations*) [Ji, 2021; Psychoula *et al*., 2021].

The objective of the current paper is to compare explainable artificial intelligence techniques in models for the detection of frauds in credit card transactions using machine learning. Specific goals in this context are: selection of datasets and classification algorithms used to detect credit card fraud, choice of explainable artificial intelligence techniques and comparison of these techniques. Apart from presenting the results, this paper also discusses the effectiveness and need

for the use of explainable artificial intelligence. In summary, the current paper contributes by presenting a robust empirical study of SHAP and LIME using three datasets, also including an evaluation of processing time.

This paper is an extended version of a previous paper [Lima and Pisani, 2024], which included part of the research and results from a work[1] presented for the Computer Science Bachelor course at Universidade Federal do ABC (UFABC). The extended version includes hyperparameter tuning, new results and an evaluation of the processing time of the explainable AI techniques. The paper is organized in the following way: in Section 2, some key concepts are introduced and related work is presented; in Section 3, the experimental methodology is presented, including datasets, hyperparameter tuning and other details from the experimental setup; in Section 4, the results are discussed; and, in Section 5, final conclusions and future work is presented.

2 Explainable AI and credit card fraud detection

Financial fraud is an important problem, in which machine learning algorithms have been applied to prevent it. There are several types of fraud, which can be categorized into internal or external. Internal fraud occurs when it is originated from individuals within a company (e.g. financial statements or money laundering), while external fraud is performed by individuals outside the company (e.g. credit cards and insurance fraud) [Aros *et al*., 2024]. Concerning the external fraud, credit card is a key field for investigation. Credit card fraud may occur in different forms, such as card theft, cloning, loss of card, purchases (which can be online or not) [Khyati Chaudhary, 2012].

Detection of credit card fraud is a complex problem. Several studies have applied machine learning algorithms to deal with it [Sulaiman *et al*., 2022; Aros *et al*., 2024]. Some algorithms frequently used in this problem are: logistic regression [Makki *et al*., 2019; Khyati Chaudhary, 2012; Pozzolo *et al*., 2015], decision tree and random forest [Makki *et al*., 2019; Khyati Chaudhary, 2012; Pozzolo *et al*., 2015], artificial neural networks and support vector machines (SVM) [Makki *et al*., 2019; Pozzolo *et al*., 2015]. A challenging issue when applying machine learning to classify credit card transactions as fraud or genuine is data imbalance, since fraudulent transactions represent a small percentage of all credit card transactions. This imbalance is illustrated in Section 3.1, where the amount of genuine transactions ranges from 0.12% to 3.50%.

There are several methods to address the data imbalance problem [Krawczyk, 2016; Thabtah *et al*., 2020]. A simple way to address it would be to simply acquire more data. Nevertheless, this is not always a possibility. Consequently, solutions based on under-sampling or over-sampling can be alternatives. Under-sampling involves removing instances from the majority class while over-sampling involves increasing the amount of instances from the minority class. Some approaches for over-sampling are *random over-sampling*, which randomly replicates minority class instances, and SMOTE (*Synthetic Minority Oversampling Technique*), which generates synthetic instances from the minority class. Other approaches are, for example, cost-sensitive learning and hybrid methods [Thabtah *et al*., 2020].

This paper focuses on fraud detection from credit card transactions. In the following sections, some concepts regarding explainable artificial intelligence (AI) are introduced, followed by a discussion on the effectiveness and need for explainable AI is presented. In the end, some related work are presented.

2.1 Explainable AI

There are many ways to define interpretability in machine learning models and several concepts which may arise from the discussion of the application of interpretability, as discussed by Lipton [2018]. Interpretability can be defined as the level of understanding which a human being has from the decisions taken by a model [Miller, 2019].

Some arguments regarding the need for interpretability are transparency and trust, as well as ethical questions. Interpretability can be a pre-requisite for trust, while a model that is transparent can be more easily interpretable. While linear models can be transparent, deep neural networks do not provide algorithmic transparency. Algorithm decisions should conform to ethical standards and interpretable models can contribute to fairness [Lipton, 2018].

Explanations of machine learning models can be dependent on the model or independent of the model. Explanations that are independent of the model can be referred to as model-agnostic interpretation techniques. These techniques are more flexible since many different machine learning algorithms can be used. On the other hand, techniques which depend on the model may lead to predictive performance loss, since the range of machine learning algorithms that can be used is reduced [Molnar, 2022]. Some aspects of a model-agnostic techniques are discussed by Ribeiro *et al*. [2016a], such as model flexibility, explanation flexibility and representation flexibility.

Model-agnostic interpretation can be divided into local and global techniques. Local explanations explain each individual prediction, while global explanations consider a more general approach, illustrating how features impact the prediction on average [Molnar, 2022]. Some common model-agnostic interpretation techniques are SHAP and LIME.

SHAP (*SHapley Additive exPlanations*) [Lundberg and Lee, 2017] is a unified approach to explain the output of machine learning models using the concept of *Shapley* values [Shapley, 2016] from the game theory. LIME (*Local interpretable model-agnostic explanations*) [Ribeiro *et al*., 2016b] is a technique which uses surrogate models. These are interpretable models trained to explain individual predictions.

Apart from LIME and SHAP, other explainable AI (XAI) techniques have been proposed in the literature. Some examples are *Integrated Gradients* [Sundararajan *et al*., 2017], *Attention maps* [Kim and Canny, 2017], RISE (*Randomized Input Sampling for Explanation*) [Petsiuk *et al*., 2018]. Some recent papers [Schwalbe and Finzel, 2024; Dwivedi *et al*.,

[1] Trabalho de Conclusão - Projeto de Graduacão em Computação (PGC): "Comparativo de algoritmos de inteligência artificial explicável no mercado financeiro" presented for the Computer Science Bachelor course at Universidade Federal do ABC (UFABC).

2023] presented taxonomies of XAI studies, illustrating how previous work could be classified according to different aspects.

2.2 Discussing explanations

Artificial intelligence has been used as decision support systems. These systems can provide recommendations to the user. However, the recommendations alone may not be enough. Interpretability of the obtained models can also be required and it may be considered a prerequisite for trusting the models [Lipton, 2018]. There is also the risk of people following a wrong recommendation blindly. Explainable AI is a tool which can provide additional support to the users by presenting explanations of the predictions/recommendations [Miller, 2023].

A key aspect when using explainable AI is the evaluation of the utility of the obtained explanations. Although SHAP and LIME techniques have been applied in several contexts, simply showing *Shapley* values from SHAP or the values returned from LIME may not be enough for people taking decisions, since the output values alone may provide a limited view.

The output provided by explainable AI techniques should target the users which will receive them [Miller *et al*., 2017]. These users may not have a high degree of knowledge of the system or the explainable AI techniques. Moreover, simply providing recommendations and explanations may not be suitable for the cognitive process involved when making decisions [Miller, 2023].

The Evaluative AI [Miller, 2023] is a new conceptual framework which can address these problems. This framework, instead of just providing information for the user, convincing to accept the model output, it provides evidence for different courses of action. The new framework may help users question the decisions made [Miller, 2023].

2.3 Related work

Several studies have applied machine learning algorithms and explainable AI to credit credit fraud detection. Some of these studies are presented in this section.

Ji [2021] evaluated the efficacy of applying LIME and SHAP to provide explanations for the results obtained from the learned models. According to the authors, the users increased their trust in the system as the explanations were provided. LIME obtained a slightly better result compared to SHAP in their study. LIME and SHAP were also studied in [Psychoula *et al*., 2021], which applied these techniques to credit card fraud detection in real time using supervised and unsupervised machine learning models. In general, the authors concluded that both SHAP and LIME are suitable to provide explanations from the learned models.

Credit card fraud detection was also studied by Hsin *et al*. [2021]. The authors argued that traditional approaches based on rules to detect frauds with credit cards are not effective. Their work proposed the use of features based on behavior and segmentation features obtained from financial expertise and characteristics from the accounts. By using these features, tree-based models, which are more easily interpretable, can obtain improved results.

Wu and Wang [2021] proposed an anomaly detection framework to detect credit card fraud. The proposed framework contained a interpretability module, which was responsible for generating explanations for the predictions. Fraud detection was performed by two deep neural networks trained in a unsupervised and adversarial way. The interpretability module involved three interpretable explainers. LIME was applied in their experiments, illustrating how input features influences the predictions.

Most studies mentioned in this section considered only a single dataset for their experiments. The current paper, on the other hand, performed an evaluation using three datasets. The evaluation on more datasets, instead of only a single dataset, allowed reporting more robust results. As described in Section 3.1, each of these datasets present different imbalance ratios between fraud and genuine instances. Although several explainable AI techniques proposals have been presented, to the best of our knowledge, most papers dealing with credit card fraud detection and explainable AI applied LIME and SHAP. Consequently, the current paper focuses on these explainable AI techniques. A summary of key differences between this paper and related work is presented in Table 1.

Reference	Explainable AI	Datasets	Evaluates processing time?
Ji [2021]	SHAP/LIME	1	no
Psychoula *et al*. [2021]	SHAP/LIME	1	yes
Hsin *et al*. [2021]	New features + Tree-based model	1	no
Wu and Wang [2021]	LIME	1	no
This paper	**SHAP/LIME**	**3**	**yes**

Table 1. A comparison between this paper and related work. The table shows, for each reference, the explainable AI technique, the number of datasets and whether processing time was evaluated.

3 Experimental methodology

The current paper presents a comparison between explainable AI techniques with machine learning models applied to credit card fraud detection. The experiments considered interpretable and black-box models. This section presents details of the evaluation methodology adopted in the experiments: datasets, training and test sets, classification algorithms, hyperparameter tuning and explainable artificial intelligence methods.

3.1 Datasets

All datasets considered in the experiments involve credit card transactions. Each transaction can be classified as fraud or genuine. In this work, three datasets were selected:

- *Kaggle fraud detection*[2]: It contains data from credit card transactions performed by European cardholders in September/2013. Overall, the dataset contains 284.807 transactions which occurred in two days. Among these transactions, only 492 were classified as fraud, which

[2] `https://www.kaggle.com/mlg-ulb/creditcardfraud`

represents 0.17% of the dataset. There are 30 input features: V1 to V28, time and amount. The features V1 to V28 were obtained using PCA (*Principal Component Analisys*). This dataset has been used by several studies, such as [Dal Pozzolo *et al*., 2014], [Pozzolo *et al*., 2015], [Dal Pozzolo *et al*., 2017], [Carcillo *et al*., 2017], [Carcillo *et al*., 2019] and [Le Borgne *et al*., 2022].

- *IEEE-CIS fraud detection*[3]: This dataset contains data from credit card transactions and information from the customers obtained from the payment company. It was used in a fraud detection competition in 2019. A high level of imbalance can be observed in this dataset, similar to the Kaggle dataset described earlier: 3.39% of the transactions are fraud in IEEE-CIS fraud detection. The dataset was originally divided into two subsets, one for training and another for testing. However, this work adopted a different methodology and only the training subset was used in the experiments, which represents around 500,000 instances. This subset was later divided into training and test during the experiments, as described in Section 3.2. This subset contains 3.50% of fraud instances. Although there are 433 input features, 19 of these features were used in our experiments. A previous work also did not use all input features from this dataset [Psychoula *et al*., 2021].
- *Credit card transaction*[4]: It is a synthetic credit card transaction dataset containing 24 million instances with 12 input features [Padhi *et al*., 2021]. Among these synthetic transactions, only around 29,000 of them are fraud, which represents 0.12%. A subset of 480,000 instances of these transactions was considered in our experiments. The subset contained the same ratio of fraud and genuine transactions.

For some datasets, there were missing data. In these cases, data imputation was applied [Moepya *et al*., 2016]. In addition, categorical features were transformed using *CatBoost* [Bourdonnaye and Daniel, 2021]. Different from the previous paper [Lima and Pisani, 2024], which applied scaling only to some input features, the experiments of the current paper applied scaling to all input features using *RobustScaler* from *scikit-learn* [Pedregosa *et al*., 2011].

3.2 Train and test sets

Training and test sets were defined using a stratified k-fold cross-validation ($k = 5$) [Alfaiz and Fati, 2022]. Consequently, five iterations were executed. In each of them, four folds were used for training and the remaining one for testing. Moreover, all experiments were executed five times, changing the random seed each time. Consequently, 25 iterations were executed (5 folds $\times$ 5 seeds).

Since there is an important data imbalance between the classes (fraud and genuine transactions), the models obtained after training could favor the classification of the majority class (genuine transactions). As a result, transactions which are fraud might not be correctly classified more frequently. To avoid this problem, we applied SMOTE [Chawla *et al*., 2002] on the datasets. Previous work also considered SMOTE when applying machine learning to detect fraud in credit card transactions [Alfaiz and Fati, 2022].

3.3 Classification algorithms

Five classification algorithms were applied in our experiments: logistic regression, decision tree, random forest, SVM and artificial neural network (*Multi-layer Perceptron*). These algorithms have been used for credit card fraud detection in previous work, as discussed in Section 2. In this paper, logistic regression, decision tree and random forest were considered interpretable algorithms, while SVM and artificial neural network were considered black-box.

Regarding the implementation, the classes *LogisticRegression*, *DecisionTreeClassifier*, *RandomForestClassifier*, *LinearSVC* and *MLPClassifier* from *scikit-learn* [Pedregosa *et al*., 2011] were applied. During the experiments, the random seed (*random_state*) from these algorithms ranged from 1 to 5. The hyperparameter configuration is described in the next section. Experiments were performed using version 1.2.2 of *scikit-learn*.

3.4 Hyperparameter tuning

An extension of the current paper compared to the previous version [Lima and Pisani, 2024] is the inclusion of hyperparameter tuning of the classification algorithms, which can significantly impact the results. Some key hyperparameters were adjusted using *Grid Search* [Bischl *et al*., 2023], while some of them were set to a specific value. *Grid search* explores exhaustively all defined hyperparameters in a systematic and reproducible way. The remaining hyperparameters assumed the default values from the scikit-learn library [Pedregosa *et al*., 2011]. *Grid search* was implemented using the *GridSearchCV*[5]. The parameter *cv* was set to a stratified k-fold cross validation ($k = 5$).

Table 2. Hyperparameter values of the classification algorithms.

Algorithm	Hyperparameter	Range of values
Logistic regression	max_iter	[2000, 5000]
	C	[0.1, 1, 10]
	solver	['newton-cholesky']
Decision tree	max_depth	[None, 5, 10]
Random forest	n_estimators	[50, 100, 200]
	max_depth	[None, 5, 10]
SVM (linear)	C	[0.1, 1, 10]
	dual	[False]
Neural network (MLP)	hidden_layer_sizes	[(100,), (50, 50)]
	max_iter	[200, 400]

Table 2 presents the hyperparameters and the range of values considered. C controls the regularization strength and is

[3] https://kaggle.com/competitions/ieee-fraud-detection
[4] https://github.com/IBM/TabFormer
[5] https://scikit-learn.org/dev/modules/generated/sklearn.model_selection.GridSearchCV.html

a critical hyperparameter for managing the model complexity. It is present in both logistic regression and SVM. In logistic regression, the maximum number of iterations (max_iter) was tuned to ensure convergence. For logistic regression, the *solver* assumed "newton-cholesky" as recommended in the *scikit-learn* documentation for datasets which contain more records than columns[6]. In addition, *dual* assumed the value "false" in *LinearSVC*, as suggested by the *scikit-learn* documentation[7] for the kind of datasets used in the experiments.

Controlling the maximum depth (max_depth) of a decision tree impacts the model complexity. Consequently, if not properly tuned, it can result in overfitting. This hyperparameter was tuned in both tree-based algorithms: decision tree and random forest. For random forest, the number of trees ($n\_estimators$) was also tuned since it can impact the prediction performance. A higher number of trees can generally improve prediction performance, but at the cost of higher processing time.

Regarding the neural network, the hyperparameter $hidden\_layer\_sizes$ defines the number of hidden layers and the number of neurons per layer. This is a key hyperparameter which directly impacts the capacity and complexity of the neural network. Additionally, the maximum number of iterations (max_iter) during training was tuned.

3.5 Explainable AI techniques

Apart from showing the predictive performance of some classification algorithms, the experiments also perform a comparison of two explainable AI techniques: SHAP and LIME.

In the experiments, SHAP was implemented using a library made available by the creators of the SHAP technique. The explainers were obtained using the class *Explainer*[8] from the SHAP library (version 0.41.0). The training data after applying SMOTE was used as *background*.

LIME was implemented using the class *LimeTabularExplainer*[9] from the *lime* library (version 0.2.0.1). The parameter *mode* assumed the value "classification", *training_data* received all training data, while the parameters *class_names* and *features_names* where adjusted for each dataset.

For both explainable AI techniques, random seed ranged from 1 to 5, similarly to Section 3.3. This parameters is known as *seed* in the SHAP library and *random_state* in the LIME library.

4 Results

The classification of credit card transactions between fraud and genuine was performed using commonly used machine learning algorithms from the literature. After the classification, explainable AI techniques SHAP and LIME were applied to assess the contribution of each input feature for all considered datasets.

This section presents and discusses the obtained results. Firstly, the overall prediction performance is shown, considering common metrics such as balanced accuracy, recall and precision. Afterwards, local explanations are presented and a comparison between the explainable AI techniques is performed. Processing time of the explainable AI techniques is also reported in a subsequent section.

4.1 Overall prediction performance

Results from some metrics are reported in Table 3. These results are the average values after five executions of the 5-fold cross validation described in Section 3.2.

A key difference of the current paper compared to the previous version [Lima and Pisani, 2024] is the inclusion of a hyperparameter tuning step, as described in Section 3.4. It improved the values of metrics in some cases, while it reduced the metric values in other cases. Regarding balanced accuracy, which represents the average performance for both classes (genuine and fraud), logistic regression obtained the best results in the first and the third datasets. Neural network obtained the best result in the second dataset, although by not a very large margin when compared to the logistic regression. These results suggest that interpretable models could be a suitable choice for the problem of credit card fraud detection. Moreover, Rudin [2019] argues that critical systems should be interpretable, avoiding the application of black-box models.

Best values for specificity and precision were obtained by random forest in all datasets. However, although sensibility and specificity values were higher than 60% in most cases, precision values were lower than 10% in several cases, sometimes lower than 1%, as reported in the third dataset. This can be explained by the very high imbalance ratio of the datasets. We will present an example considering the third dataset, which has only 0.12% of fraud transactions: 479402 genuine transaction and 598 fraud transactions. Considering an hypothetical scenario where sensibility is 100% (all frauds were detected) and specificity is 99%, the precision would be around 11% ($0.11 = 0.0012/(0.9988.(0.01) + 0.0012)$[10]. In the same hypothetical scenario, a lower specificity of 95% would lead to a precision of 2.35% ($0.0235 = 0.0012/(0.9988.(0.05) + 0.0012)$). In summary, due to high imbalance ratio, it is challenging to obtain a higher precision value. The low precision values suggest a high amount of false alarms, but, at the same time, the sensibility (rate of detected frauds) is above 75% in several cases.

4.2 Local explanations

This section presents a comparison between explainable AI techniques SHAP and LIME. First, the weights obtained af-

[6] `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html`

[7] `https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html`

[8] `https://shap.readthedocs.io/en/latest/generated/shap.Explainer.html`

[9] `https://lime-ml.readthedocs.io/en/latest/lime.html`

[10] Precision is defined as TP / (FP + TP), where TP is the amount of true positives and FP is the amount of false positives. The positive class in this paper is *fraud*, while the negative class is *genuine*. If the dataset has X instances and the amount of fraud transactions is 0.12%, there are $0.0012.X$ fraud transactions and $0.9988.X$ genuine transactions. In the example, 100% sensibility means $0.0012.X$ fraud transactions correctly classified (TP) and 99% of specificity means 1% of genuine transactions wrongly classified as fraud (FP): $0.9988.X.(0.01)$. Hence, precision = $(0.0012.X)/(0.9988.(0.01).X + 0.0012.X) = 0.0012/(0.9988.(0.01) + 0.0012) = 0.11$.

Kaggle fraud detection					
	Decision tree	**Logistic regression**	**Random forest**	**SVM**	**Neural network**
Balanced accuracy	0.8876 (0.0179)	**0.9427 (0.0136)**	0.9113 (0.0167)	0.9414 (0.0156)	0.9018 (0.0212)
Sensibility	0.7776 (0.0358)	**0.9093 (0.0282)**	0.8227 (0.0334)	0.9024 (0.0318)	0.8040 (0.0424)
Specificity	0.9977 (0.0002)	0.9762 (0.0019)	**0.9998 (0.0001)**	0.9805 (0.0016)	0.9995 (0.0002)
Precision	0.3715 (0.0232)	0.0622 (0.0037)	**0.8935 (0.0379)**	0.0745 (0.0049)	0.7564 (0.0630)
IEEE-CIS fraud detection					
	Decision tree	**Logistic regression**	**Random forest**	**SVM**	**Neural network**
Balanced accuracy	0.5924 (0.0225)	0.6920 (0.0021)	0.6075 (0.0265)	0.6912 (0.0022)	**0.7198 (0.0141)**
Sensibility	0.3790 (0.0274)	**0.6294 (0.0162)**	0.3045 (0.0534)	0.6232 (0.0158)	0.6209 (0.0507)
Specificity	0.8058 (0.0562)	0.7547 (0.0158)	**0.9104 (0.0624)**	0.7592 (0.0153)	0.8186 (0.0414)
Precision	0.0698 (0.0155)	0.0853 (0.0032)	**0.1586 (0.0952)**	0.0860 (0.0032)	0.1136 (0.0161)
Credit card transaction					
	Decision tree	**Logistic regression**	**Random forest**	**SVM**	**Neural network**
Balanced accuracy	0.6547 (0.0350)	**0.8331 (0.0213)**	0.6385 (0.0413)	0.8320 (0.0226)	0.7169 (0.0271)
Sensibility	0.3264 (0.0718)	**0.7807 (0.0539)**	0.2782 (0.0831)	0.7720 (0.0555)	0.4435 (0.0540)
Specificity	0.9829 (0.0062)	0.8856 (0.0154)	**0.9988 (0.0009)**	0.8921 (0.0137)	0.9902 (0.0022)
Precision	0.0256 (0.0085)	0.0086 (0.0013)	**0.2950 (0.1601)**	0.0090 (0.0012)	0.0560 (0.0134)

Table 3. Predictive performance of the machine learning models in each of the datasets. The best result for each metric per dataset are highlighted in bold. Standard deviation is reported between parenthesis.

	Kaggle Fraud Detection		**IEEE-CIS Fraud Detection**		**Credit Card Transaction**	
	SHAP	**LIME**	**SHAP**	**LIME**	**SHAP**	**LIME**
Decision tree	**56%**	55%	39%	**45%**	**84%**	82%
SVM	**71%**	67%	**66%**	58%	**87%**	82%
Logistic regression	**74%**	66%	**65%**	59%	**87%**	81%
Neural network	52%	**54%**	**49%**	40%	**84%**	82%
Random forest	64%	**66%**	40%	**43%**	**84%**	82%

Table 4. Comparison of SHAP and LIME techniques in each dataset. The evaluation considered the number of times that each of the 10 most important input features from LIME and SHAP occurred in the list of the most important input features obtained by the logistic regression weights in their respectively cross-validation iteration (the order in which each feature occurred in the lists was not considered). The best results are highlighted in bold.

ter fitting a logistic regression were used to obtain a list of the most important input features in each dataset, similarly to the ranking discussed in [Psychoula *et al*., 2021]. This list was used as a reference to evaluate the results of SHAP and LIME when providing the explanations from the selected instances. The adopted methodology corresponds to a functionally-grounded evaluation [Doshi-Velez and Kim, 2017], in which the logistic regression weights were defined as a *proxy*. Logistic regression is a suitable choice to obtain feature weights due to the transparency of the models and their capacity to meet regulatory requirements [Psychoula *et al*., 2021; Bücker *et al*., 2022].

The 10 most important input features, according to the logistic regression weights, for each dataset are shown in Figure 1. In the first dataset, most input features were obtained after a transformation using PCA as described in Section 3.1. The use of this transformation makes it difficult to obtain explanations from users without knowledge of the codification process. However, the evaluation of explainable AI techniques in this case is still important to check their performance in transformed datasets, which is a common situation when dealing with datasets from credit card fraud detection [Psychoula *et al*., 2021].

Once the 10 most important predictors were defined by the logistic regression weights, the evaluation of the explainable AI was performed. The evaluation considered local explanations for some instances from the positive class (fraud) in the training dataset. For each iteration of the 5-fold stratified cross-validation, one instance was randomly selected. Explainer models were obtained using training data, since the goal was to study the most important input features for each evaluated model [Molnar, 2022]. From the obtained models, the most important input features were identified. The importance of the features for the evaluation considered absolute *Shapley* values obtained by SHAP and absolute values output by LIME. The results for SHAP in each dataset are shown in Figures 2, 3, 4, while the results for LIME are presented in Figures 5, 6, 7. The results on these figures are from the first random seed.

From the lists of most important input features, the comparison of SHAP and LIME techniques was performed. The comparison consisted of computing the amount of input features from the top-10 list of each explainable AI technique which were among the top-10 list of the logistic regression weights. The order in which each input feature occurred was not considered for this evaluation. Since five instances were considered (5-fold) and, for each of them, the maximum amount of common input features is 10, the maximum value of the sum is 50 per random seed (the maximum sum is 250 considering all five random seeds). A summary of this evalu-

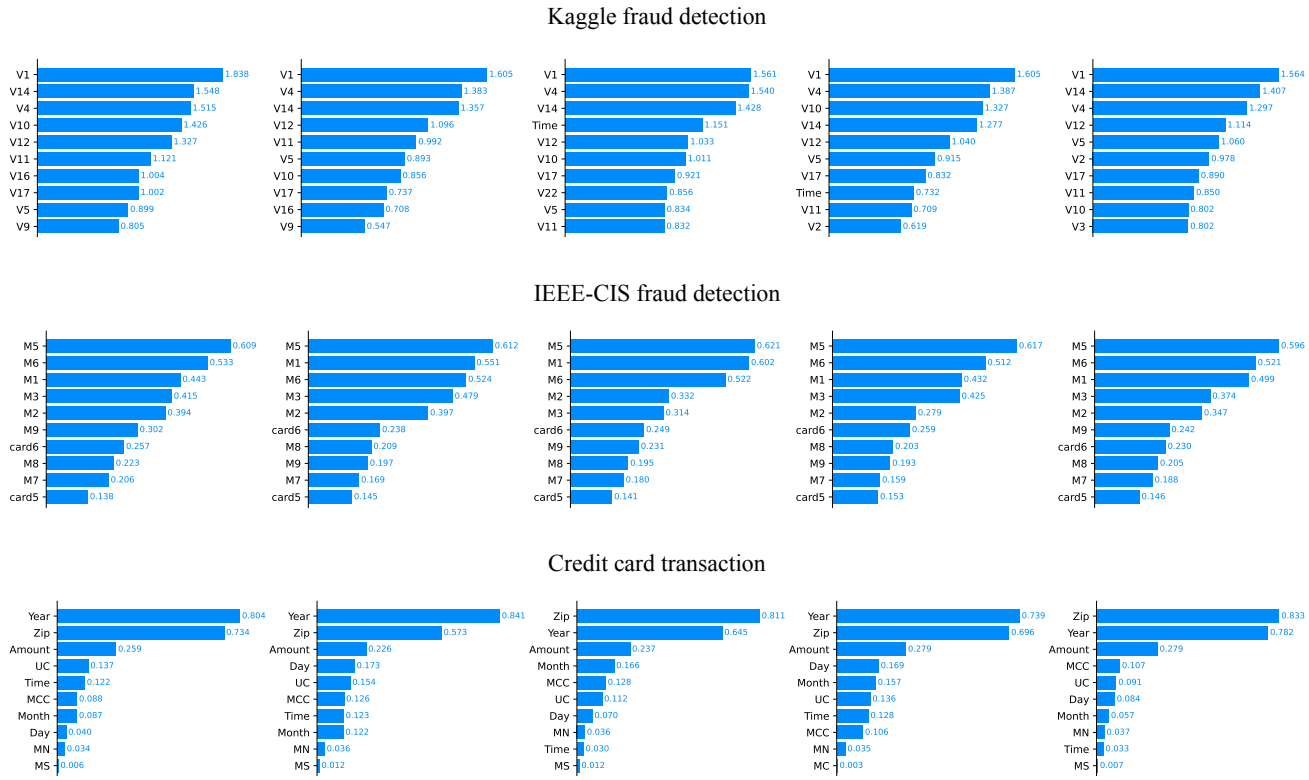

Figure 1. List of the 10 most important features according to the absolute value of the weights of the logistic regression models (globally computed). There are five plots per dataset, one for each iteration of the 5-fold cross-validation.

ation is shown in Table 4. The reported values are the overall number of common features (in percentage) considering each iteration of the cross-validation for all five random seeds.

The results reported in Table 4 show that, overall, the performance of SHAP was higher than LIME, though LIME obtained better results in some cases. Another aspect that can be observed is the variation in the order of the top-10 input features in the 5-fold cross-validation, which changes training and test sets at each iteration. This variation is illustrated in Figures 2, 3, 4, 5, 6 and 7. The values of SHAP and LIME changed among folds and the order of the top-10 features did not remain the same for all instances. It suggests that these techniques are susceptible to perturbations [Alvarez-Melis and Jaakkola, 2018].

4.3 Processing time to obtain explanations

In addition to the prediction performance and the comparison of the explainable AI techniques, this paper includes an evaluation of the processing time to obtain SHAP and LIME values. The results are presented in Table 5 (SHAP) and in Table 6 (LIME). All experiments were executed in the same computer (Intel Core i7-8550U, 16 GB of RAM).

The LIME library requires the creation of an explainer, which consumes more processing time than SHAP. Although it may seem as a drawback from LIME compared to SHAP, this explainer can be used for all explanations in the dataset afterwards. Regarding the time to output the explanations (SHAP and LIME values), LIME was faster in most cases. It indicates that, although creating the LIME explainer can consume more computer resources than SHAP, the explanations

Kaggle fraud detection	
	SHAP Values
Logistic regression	0.11s
Decision tree	0.02s
Random forest	0.49s
SVM	0.01s
Neural network	0.04s
IEEE-CIS fraud detection	
	SHAP Values
Logistic regression	0.11s
Decision tree	0.02s
Random forest	1.72s
SVM	0.02s
Neural network	0.03s
Credit card transaction	
	SHAP Values
Logistic regression	0.08s
Decision tree	0.03s
Random forest	1.03s
SVM	0.03s
Neural network	0.06s

Table 5. Processing time to obtain SHAP values.

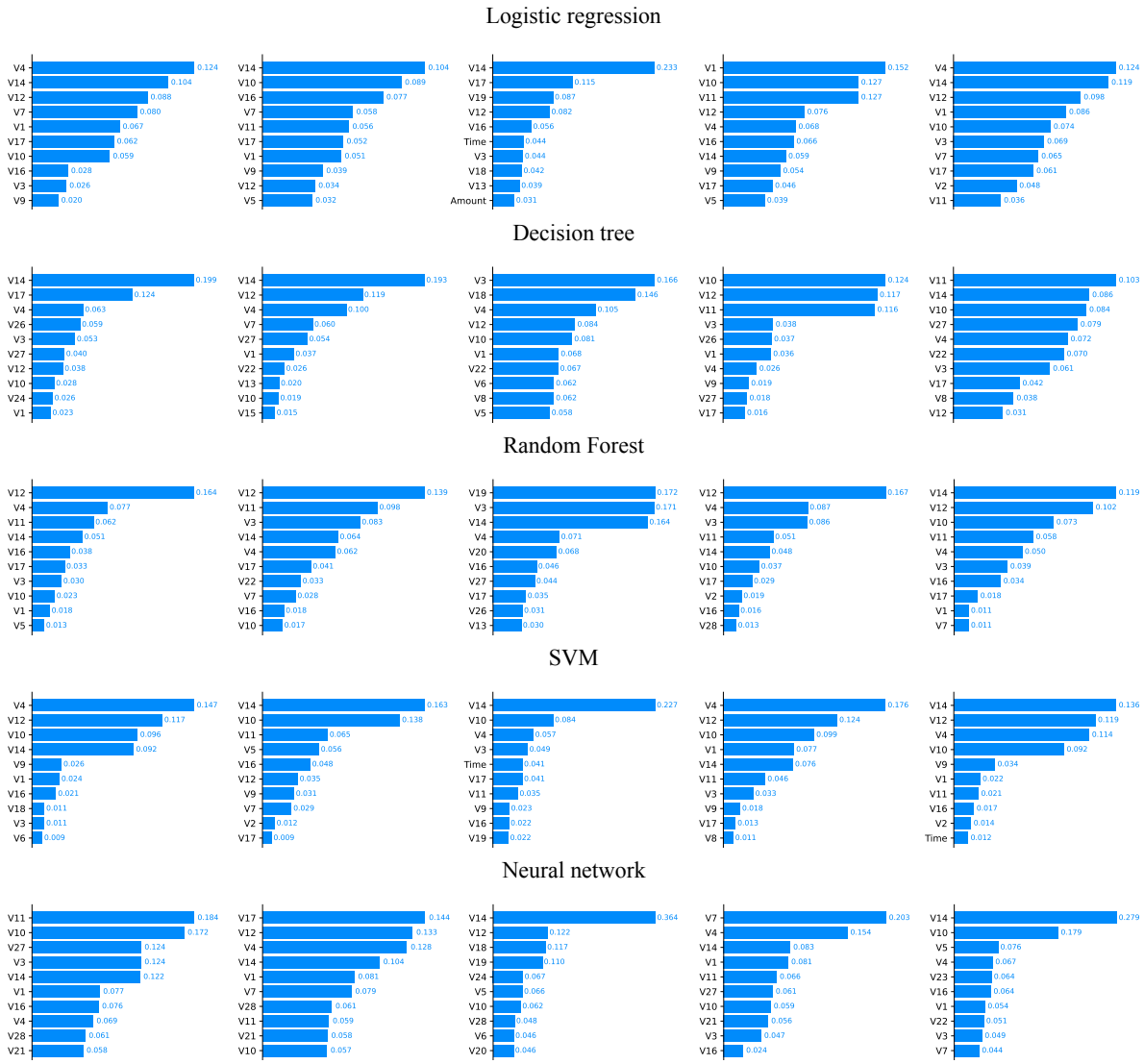

Figure 2. Top 10 features by absolute SHAP value for each algorithm, evaluated on the *Kaggle fraud detection* dataset (first random seed).

Kaggle fraud detection		
	LIME Explainer	**LIME Values**
Logistic regression	3.41s	0.05s
Decision tree	3.46s	0.04s
Random forest	3.25s	0.12s
SVM	3.26s	0.05s
Neural network	3.90s	0.06s
IEEE-CIS fraud detection		
	LIME Explainer	**LIME Values**
Logistic regression	4.15s	0.03s
Decision tree	4.03s	0.03s
Random forest	4.22s	0.41s
SVM	4.02s	0.03s
Neural network	4.03s	0.03s
Credit card transaction		
	LIME Explainer	**LIME Values**
Logistic regression	1.98s	0.02s
Decision tree	1.91s	0.02s
Random forest	1.92s	0.10s
SVM	1.91s	0.02s
Neural network	1.98s	0.02s

Table 6. Processing time to obtain LIME explainer and values.

obtained afterwards can usually be obtained much faster by LIME.

5 Conclusion

This paper presented a comparison of the application of explainable AI techniques SHAP and LIME in the context of credit card fraud detection. During the experiments, these techniques were applied to five classification algorithms in three datasets. A key aspect present in all datasets is the high degree of imbalance, which represents an important challenge to the task of classifying the credit card transactions between fraud and genuine. The current work is an extended version of a previous study [Lima and Pisani, 2024] and includes hyperparameter tuning, new results and an evaluation of the processing time of the explainable AI techniques.

Firstly, an evaluation of the overall prediction performance was performed. The best balanced accuracy was obtained by logistic regression in two datasets. Neural network obtained the best value in another dataset. although not by a very large margin when compared to the logistic regression. It suggests that interpretable algorithms can be considered for credit card

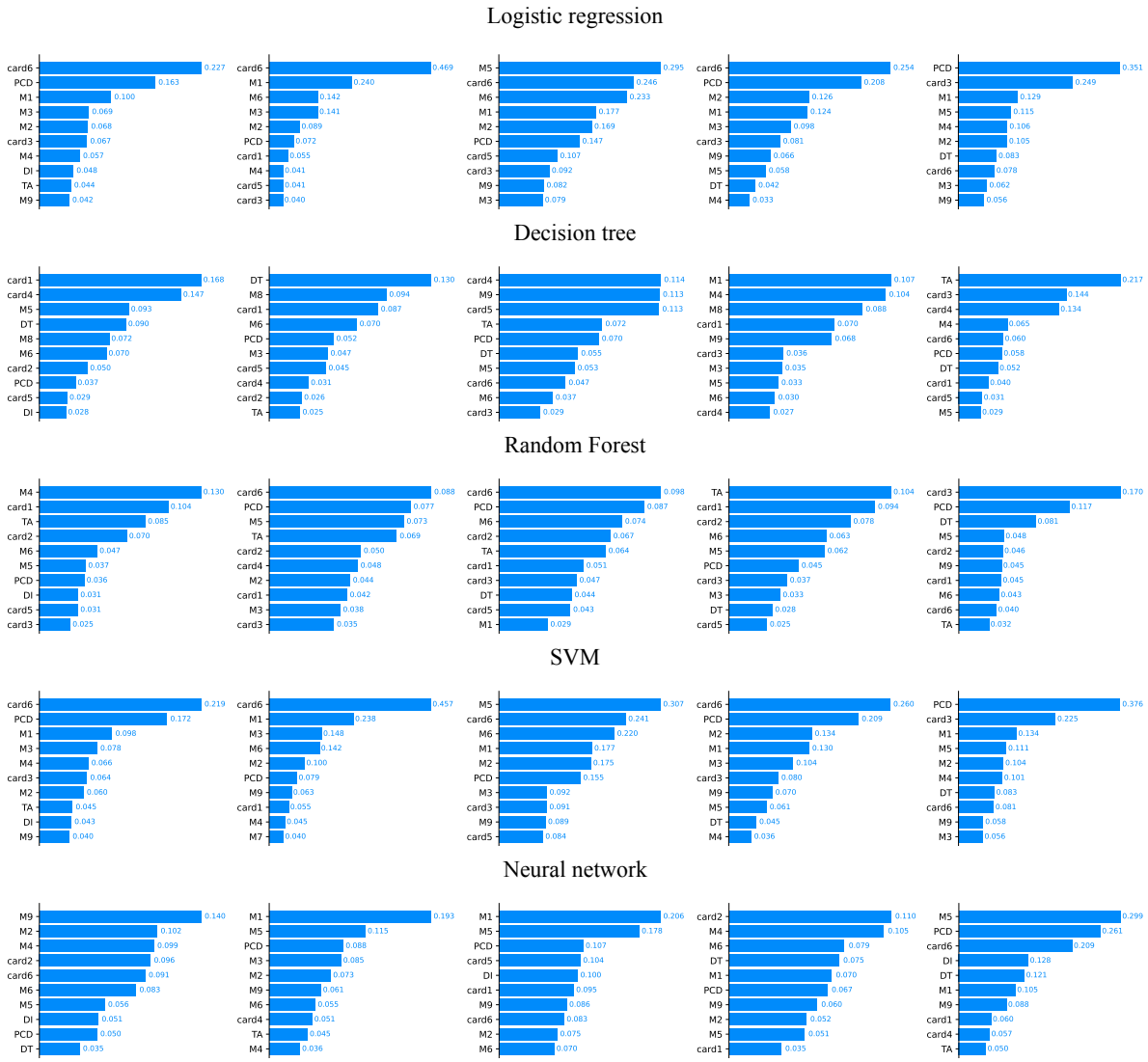

Figure 3. Top 10 features by absolute SHAP value for each algorithm, evaluated on the *IEEE-CIS fraud detection* dataset (first random seed).

fraud detection and could contribute to the argument from Rudin [2019] that critical systems should use interpretable models.

An evaluation of the explainable AI techniques was presented after the initial evaluation of the prediction performance. Using logistic regression models as the proxy, the evaluation was based on the 10 most important input features according to each technique. Overall, SHAP obtained better results than LIME in our experiments. The current paper also evaluated the processing time of the explainable AI techniques. In the experiments, SHAP consumed more processing time than LIME to obtain the explanations in most cases, although LIME requires additional time for the explainer.

In this work, only local explanations were obtained. Future work could explore global explanation and check how it compares to the global weights of the logistic regression. The evaluation of global explanations could also potentially reduce bias due to specific random choice of an instance from the dataset. A study regarding the representativeness of the training and test sets may also be carried out in the future.

The evaluation of the quality of explanations can be considered an open question in the literature since the definition of interpretability lacks mathematical rigor [Aldeia and de França, 2022; Marcinkevičs and Vogt, 2023; Lipton, 2018]. However, some metrics for evaluating explainable AI methods have been discussed in the literature and could be considered in future work [Schwalbe and Finzel, 2024]. Moreover, a more comprehensive study of the quality of the interpretations would require the participation of humans to provide feedback regarding the explainable AI techniques, similar to the approach described in [Doshi-Velez and Kim, 2017].

Studying methods to report results is another topic for future research in the context of credit card fraud detection. Since there are sensitive information in the datasets in this context, which are frequently transformed, it is a challenge obtaining useful and reliable explanations. The same discussion applies to the models. In the context of the problem, the actual contribution of a transformed input feature may be hard to define. A unified evaluation of all datasets considering common input features is also a topic for future work.

Regarding dataset size, both factors, number of transactions and number of features can impact the performance. In future work, a study of how the dataset size impacts the processing time could be performed, controlling the number of transactions and features.

Security aspects of the explainable AI techniques is another

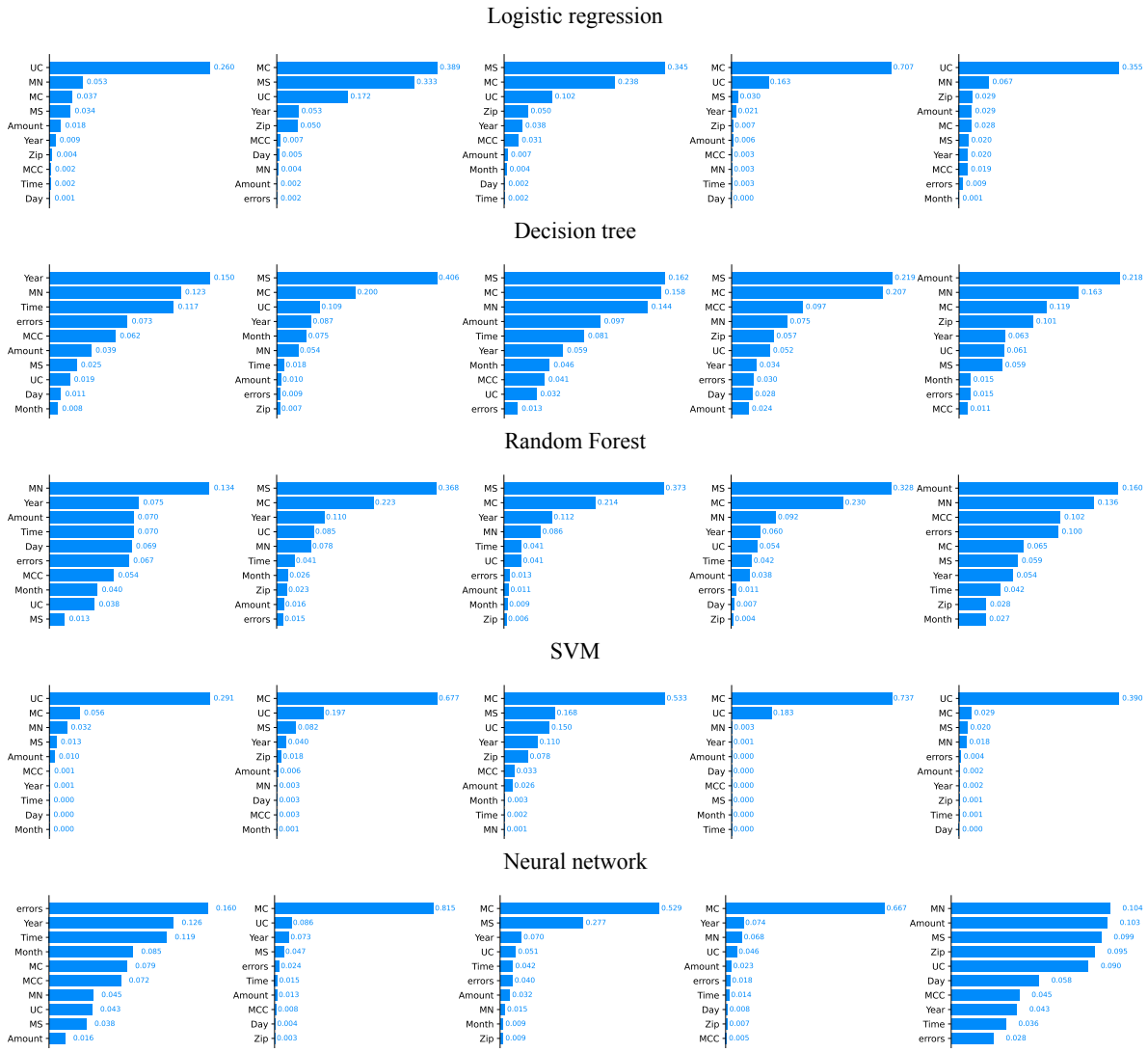

Figure 4. Top 10 features by absolute SHAP value for each algorithm, evaluated on the *Credit card transaction* dataset (first random seed).

alternative for additional research. Explainable AI has been employed in several CyberSecurity applications, including fraud detection, and there are open challenges which can be addressed in future work [Capuano *et al.*, 2022].

Declarations

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

Authors' Contributions

GL and PP contributed to the conception of this study. GL implemented the source code and performed the experiments. GL reported all results from the experiments, including tables and plots. PP and GL wrote the paper. All authors read and approved the final manuscript.

Competing interests

The authors declare that they do not have any competing interests.

Availability of data and materials

The source code of the experiments performed for this paper is available at `https://github.com/GaMendes/jbcs-credit-card-fraud`.

References

Aldeia, G. S. I. and de França, F. O. (2022). Interpretability in symbolic regression: a benchmark of explanatory methods using the feynman data set. *Genetic Programming and Evolvable Machines*, 23:309–349. DOI: 10.1007/s10710-022-09435-x.

Alfaiz, N. S. and Fati, S. M. (2022). Enhanced credit card fraud detection model using machine learning. *Electronics (Switzerland)*, 11. DOI: 10.3390/electronics11040662.

Alvarez-Melis, D. and Jaakkola, T. S. (2018). On the robustness of interpretability methods. DOI: 10.48550/arxiv.1806.08049.

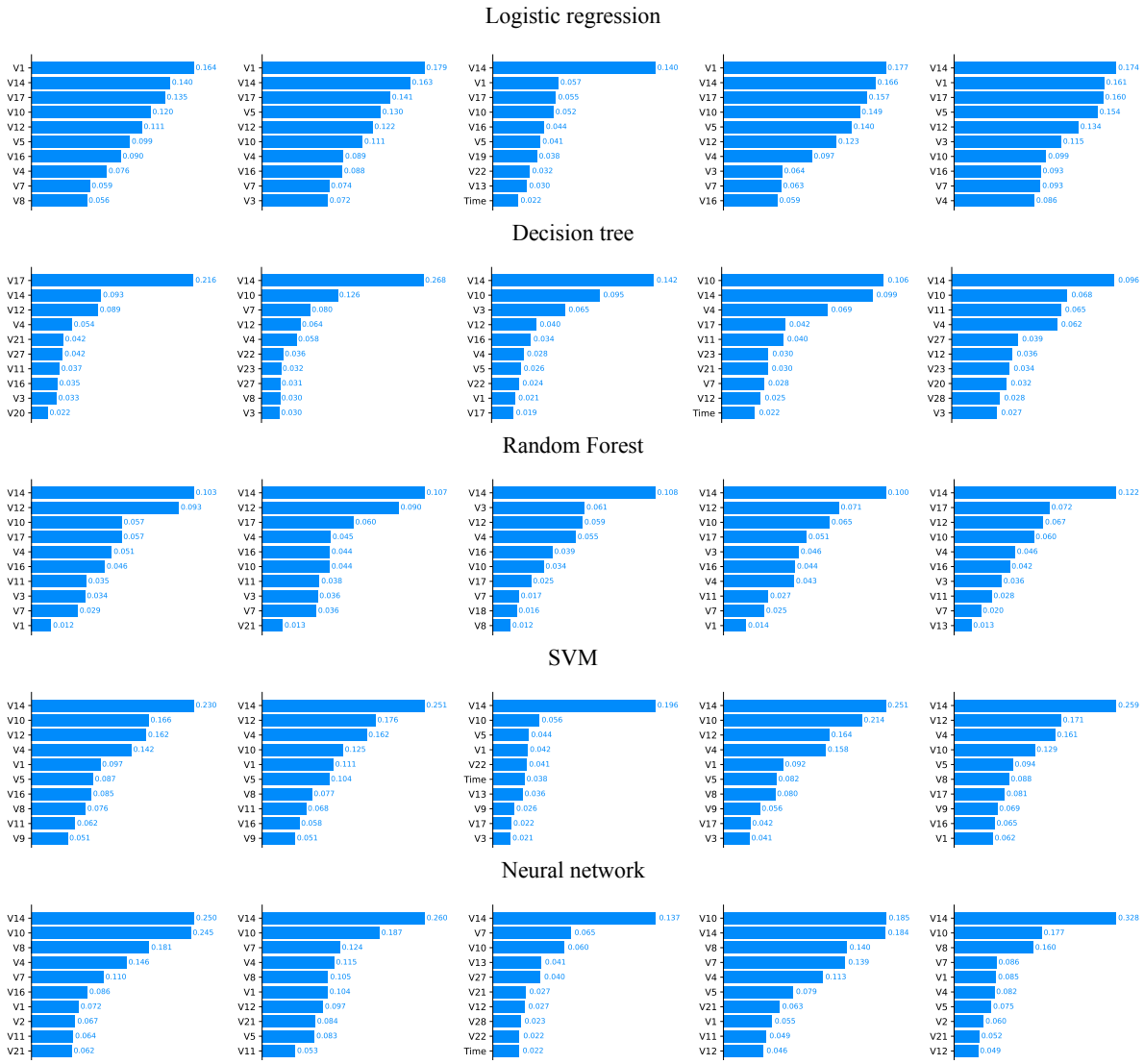

Figure 5. Top 10 features by absolute LIME value for each algorithm, evaluated on the *Kaggle fraud detection* dataset (first random seed).

Aros, L. H., Molano, L. X. B., Gutierrez-Portela, F., Hernandez, J. J. M., and Barrero, M. S. R. (2024). Financial fraud detection through the application of machine learning techniques: a literature review. *Humanities and Social Sciences Communications*. DOI: 10.1057/s41599-024-03606-0.

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.-L., Deng, D., and Lindauer, M. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *WIREs Data Mining and Knowledge Discovery*, 13(2):e1484. DOI: 10.1002/widm.1484.

Bourdonnaye, F. D. L. and Daniel, F. (2021). Evaluating categorical encoding methods on a real credit card fraud detection database. Available at:://arxiv.org/pdf/2112.12024.

Bussmann, N., Giudici, P., Marinelli, D., and Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics*, 57:203–216. DOI: 10.1007/s10614-020-10042-0.

Bücker, M., Szepannek, G., Gosiewska, A., and Biecek, P. (2022). Transparency, auditability, and explainability of machine learning models in credit scoring. *Journal of the Operational Research Society*, 73(1):70–90. DOI: 10.1080/01605682.2021.1922098.

Capuano, N., Fenza, G., Loia, V., and Stanzione, C. (2022). Explainable artificial intelligence in cybersecurity: A survey. *IEEE Access*, 10:93575–93600. DOI: 10.1109/ACCESS.2022.3204171.

Carcillo, F., Dal Pozzolo, A., Le Borgne, Y.-A., Caelen, O., Mazzer, Y., and Bontempi, G. (2017). Scarff : a scalable framework for streaming credit card fraud detection with spark. *Information Fusion*, 41. DOI: 10.1016/j.inffus.2017.09.005.

Carcillo, F., Le Borgne, Y.-A., Caelen, O., Kessaci, Y., Oblé, F., and Bontempi, G. (2019). Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences*. DOI: 10.1016/j.ins.2019.05.042.

Černevičienė, J. and Kabašinskas, A. (2024). Explainable artificial intelligence (xai) in finance: a systematic literature review. *Artificial Intelligence Review*, 57(8):216. DOI: 10.1007/s10462-024-10854-8.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357. DOI: 10.1613/jair.953.

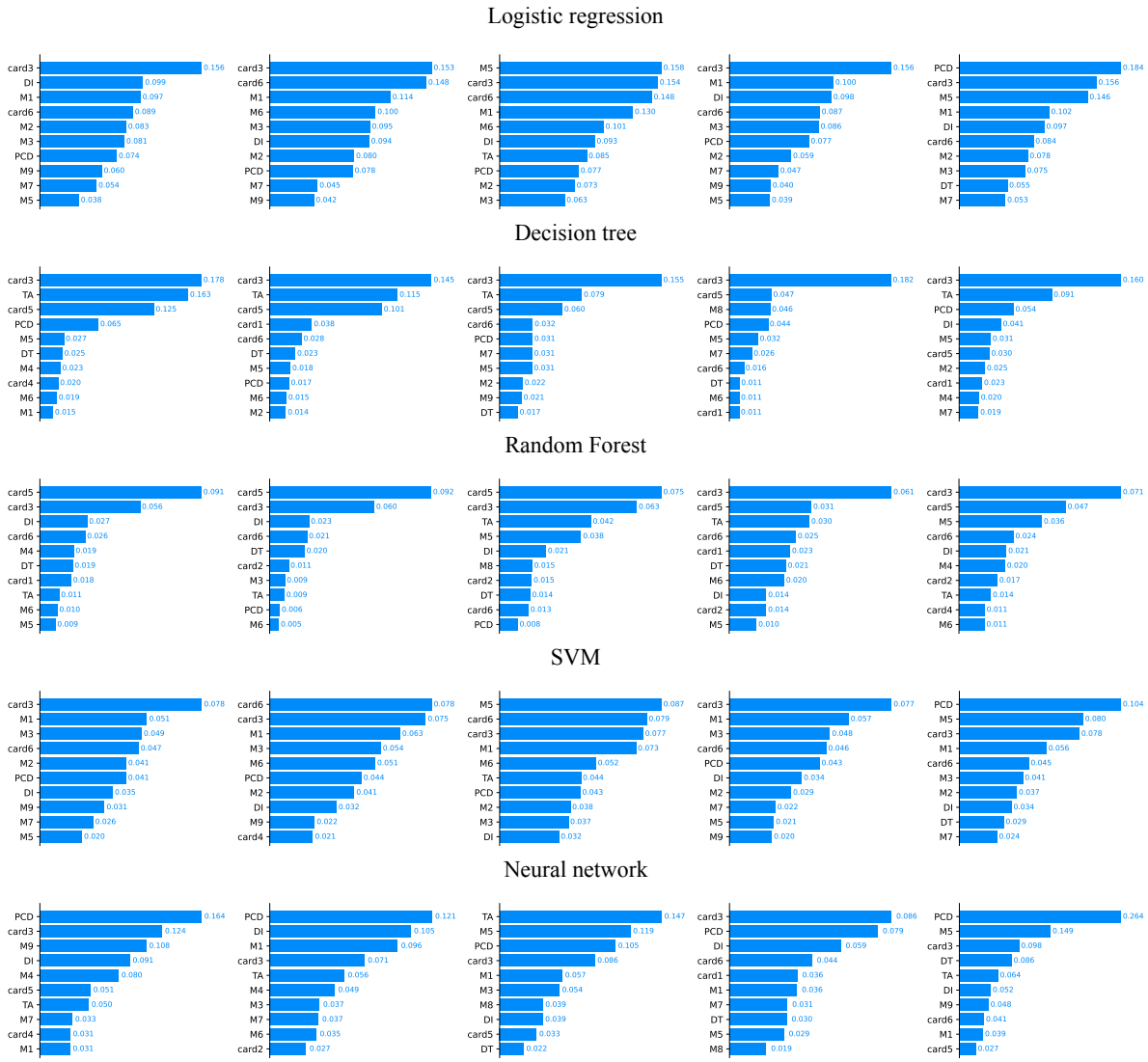

Figure 6. Top 10 features by absolute LIME value for each algorithm, evaluated on the *IEEE-CIS fraud detection* dataset (first random seed).

Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., and Bontempi, G. (2017). Credit card fraud detection: A realistic modeling and a novel learning strategy. *IEEE Transactions on Neural Networks and Learning Systems*, PP:1–14. DOI: 10.1109/TNNLS.2017.2736643.

Dal Pozzolo, A., Caelen, O., Le Borgne, Y.-A., Waterschoot, S., and Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications*, 41:4915–4928. DOI: 10.1016/j.eswa.2014.02.026.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. DOI: 10.48550/arxiv.1702.08608.

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., and Ranjan, R. (2023). Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Comput. Surv.*, 55(9). DOI: 10.1145/3561048.

European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal L110*, 59:1–88. Available at:`https://eur-lex.europa.eu/eli/reg/2016/679/oj`.

Gee, J., Button, M., and Brooks, G. (2019). *The financial cost of fraud: what data from around the world shows*. MacIntyre Hudson. Book.

Hanif, A. (2021). Towards explainable artificial intelligence in banking and financial services. DOI: 10.48550/arxiv.2112.08441.

Hsin, Y.-Y., Dai, T.-S., Ti, Y.-W., and Huang, M.-C. (2021). Interpretable electronic transfer fraud detection with expert feature constructions. In *CIKM Workshops*. Avaialble at:`https://scholar.nycu.edu.tw/en/publications/interpretable-electronic-transfer-fraud-detection-wit`

Ji, Y. (2021). Explainable ai methods for credit card fraud detection: Evaluation of LIME and SHAP through a user study. Available at:`https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1626230`.

Khyati Chaudhary, Jyoti Yadav, B. M. (2012). A review of fraud detection techniques: Credit card. *International Journal of Computer Applications*, 45(1):39–44. DOI: 10.5120/6748-8991.

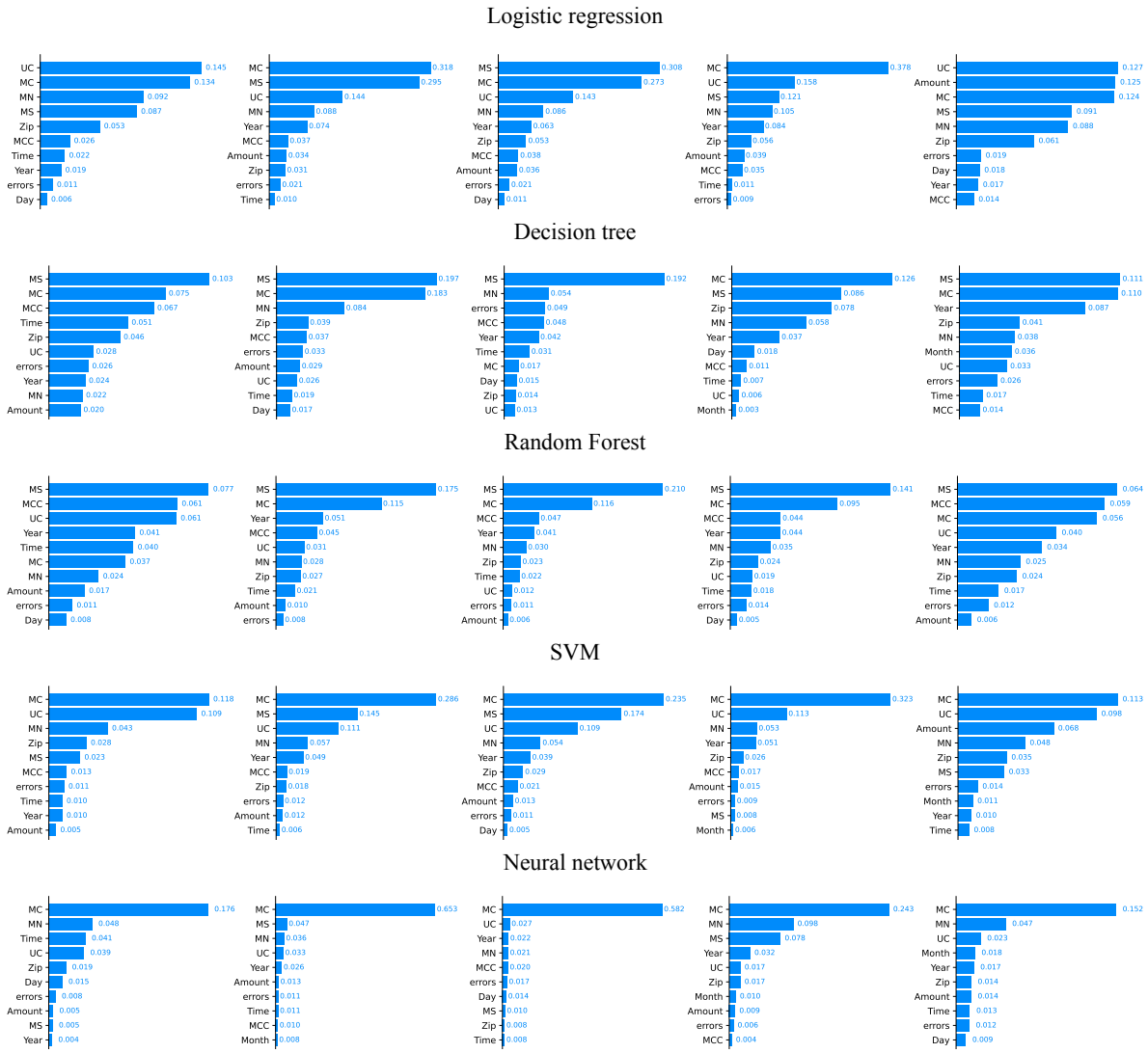

Figure 7. Top 10 features by absolute LIME value for each algorithm, evaluated on the *Credit card transaction* dataset (first random seed).

Kim, J. and Canny, J. (2017). Interpretable learning for self-driving cars by visualizing causal attention. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969. DOI: 10.1109/ICCV.2017.320.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5:221–232. DOI: 10.1007/s13748-016-0094-0.

Le Borgne, Y.-A., Siblini, W., Lebichot, B., and Bontempi, G. (2022). *Reproducible Machine Learning for Credit Card Fraud Detection - Practical Handbook*. Université Libre de Bruxelles. Available at:`https://github.com/Fraud-Detection-Handbook/fraud-detection-handbook`.

Lima, G. M. d. and Pisani, P. H. (2024). Comparativo de técnicas de inteligência artificial explicável na detecção de fraudes em transações com cartão de crédito. In *Anais Estendidos do XXIV Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais*, pages 244–255, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbseg_estendido.2024.243180.

Lipton, Z. C. (2018). The mythos of model interpretability. *Commun. ACM*, 61(10):36–43. DOI: 10.1145/3233231.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.. DOI: 10.48550/arxiv.1705.07874.

Makki, S., Assaghir, Z., Taher, Y., Haque, R., Hacid, M. S., and Zeineddine, H. (2019). An experimental study with imbalanced classification approaches for credit card fraud detection. *IEEE Access*, 7:93010–93022. DOI: 10.1109/ACCESS.2019.2927266.

Marcinkevičs, R. and Vogt, J. E. (2023). Interpretability and explainability: A machine learning zoo mini-tour. DOI: 10.48550/arxiv.2012.01805.

Martins, T., de Almeida, A. M., Cardoso, E., and Nunes, L. (2024). Explainable artificial intelligence (xai): A systematic literature review on taxonomies and applications in finance. *IEEE Access*, 12:618–629. DOI: 10.1109/ACCESS.2023.3347028.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38. DOI: 10.1016/j.artint.2018.07.007.

Miller, T. (2023). Explainable AI is Dead, Long Live Explain-

able AI! Hypothesis-driven decision support. Available at:https://arxiv.org/pdf/2302.12389.

Miller, T., Howe, P., and Sonenberg, L. (2017). Explainable ai: Beware of inmates running the asylum. Available at:https://arxiv.org/pdf/1712.00547.

Moepya, S. O., Akhoury, S. S., Nelwamondo, F. V., and Twala, B. (2016). The role of imputation in detecting fraudulent financial reporting. *International Journal of Innovative Computing, Information and Control ICIC International c*, 12:333–356. Available at:https://www.researchgate.net/publication/297047624_The_role_of_imputation_in_detecting_fraudulent_financial_reporting.

Molnar, C. (2022). *Interpretable Machine Learning*. 2 edition. Available at:https://christophm.github.io/interpretable-ml-book.

Padhi, I., Schiff, Y., Melnyk, I., Rigotti, M., Mroueh, Y., Dognin, P., Ross, J., Nair, R., and Altman, E. (2021). Tabular transformers for modeling multivariate time series. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3565–3569. DOI: 10.1109/ICASSP39728.2021.9414142.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. Avaialble at:https://jmlr.org/papers/v12/pedregosa11a.html.

Petsiuk, V., Das, A., and Saenko, K. (2018). Rise: randomized input sampling for explanation of black-box models. In *Proceedings of the British machine vision conference*, pages 1–13. DOI: 10.48550/arxiv.1806.07421.

Pozzolo, A. D., Caelen, O., Johnson, R. A., and Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE Symposium Series on Computational Intelligence*, pages 159–166. DOI: 10.1109/SSCI.2015.33.

Psychoula, I., Gutmann, A., Mainali, P., Lee, S. H., Dunphy, P., and Petitcolas, F. (2021). Explainable machine learning for fraud detection. *Computer*, 54(10):49–59. DOI: 10.1109/MC.2021.3081249.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016a). Model-agnostic interpretability of machine learning. pages 91–95. DOI: 10.48550/arxiv.1606.05386.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016b). "Why should I trust you?" explaining the predictions of any classifier. volume 13-17-August-2016, pages 1135–1144. Association for Computing Machinery. DOI: 10.1145/2939672.2939778.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215. DOI: 10.1038/s42256-019-0048-x.

Schwalbe, G. and Finzel, B. (2024). A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, 38:3043–3101. DOI: 10.1007/s10618-022-00867-8.

Shapley, L. S. (2016). *17. A Value for n-Person Games*, pages 307–318. Princeton University Press, Princeton. DOI: 10.1515/9781400881970-018.

Sulaiman, R. B., Schetinin, V., and Sant, P. (2022). Review of machine learning approach on credit card fraud detection. *Human-Centric Intelligent Systems*, pages 55–68. DOI: 10.1007/s44230-022-00004-0.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR. DOI: 10.48550/arxiv.1703.01365.

Thabtah, F., Hammoud, S., Kamalov, F., and Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. *Information Sciences*, 513:429–441. DOI: 10.1016/j.ins.2019.11.004.

Wu, T.-Y. and Wang, Y.-T. (2021). Locally interpretable one-class anomaly detection for credit card fraud detection. In *2021 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pages 25–30. DOI: 10.1109/TAAI54685.2021.00014.