Evaluation of XAI Methods in a FinTech Context

Falko Gawantka[1(✉)], Franz Just[2], Markus Ullrich[1], Marina Savelyeva[1], and Jörg Lässig[1,3]

[1] University of Applied Sciences Zittau/Görlitz, 02826 Görlitz, Germany
{Falko.Gawantka,M.Ullrich,Marina.Savelyeva}@hszg.de
[2] Universidad de Granada, 18012 Granada, Spain
Franz.Just@hszg.de
[3] Fraunhofer IOSB-AST, 02826 Görlitz, Germany
joerg.laessig@iosb-ast.fraunhofer.de

Abstract. As humans, we automate more and more critical areas of our lives while using machine learning algorithms to make autonomous decisions. For example, these algorithms may approve or reject job applications/loans. To ensure the fairness and reliability of the decision-making process, a validation is required. The solution for explaining the decision process of ML models is Explainable Artificial Intelligence (XAI). In this paper, we evaluate four different XAI approaches - LIME, SHAP, CIU, and Integrated Gradients (IG) - based on the similarity of their explanations. We compare their feature importance values (FIV) and rank the approaches from the most trustworthy to the least trustworthy. This ranking can serve as a specific fidelity measure of the explanations provided by the XAI methods.

Keywords: ML · XAI · Local Interpretable Model-Agnostic Explanations (LIME) · SHapley Additive exPlanations (SHAP) · Contextual Importance and Utility (CIU) · Integrated Gradients (IG)

1 Introduction and Motivation

Automated systems are increasingly present in various aspects of our lives. Current research shows that it is even possible to use AI to automate the processing of job applications so that positions can be filled as quickly as possible and suitable candidates can be found more efficiently [1]. In the field of medical imaging, AI is being used as a decision support system to more effectively evaluate the large amounts of data generated by procedures such as MRI [13,19]. AI also has a significant impact on the financial sector, with the growing field of financial technology (FinTech) using AI in decision-making processes, including lending and insurance [3,12]. However, there is a gap in providing meaningful information to human decision makers. Given the vast amounts of data, these decision-makers rely on AI evaluations, but these evaluations should also be justified.

This work is supported with tax funds on the basis of the budget passed by the Saxon State Parliament.

Y. Hernndez Heredia et al. (Eds.): IWAIPR 2023, LNCS 14335, pp. 143–154, 2024.
https://doi.org/10.1007/978-3-031-49552-6_13

2 Problem Description and Motivation

In the financial context, automated decisions can have a critical impact on individuals. Credit approval serves as a representative use case to illustrate the limitations and issues of using AI predictions in this domain. To address the lack of transparency, Fig. 1 illustrates the AI-assisted part of the credit approval process, which relies on a predictive model built using deep learning. The model uses various features to determine an individual's ability to repay the loan.

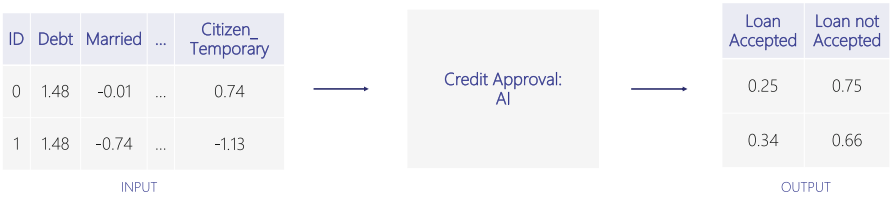

Fig. 1. An example for the lack of transparency in the credit approval use case. The input and output behavior of a black box predictor is shown. For the clarity of the *standardised* input data, there are two test instances presented and the resulting probabilities by the AI model.

The input to the *Credit Approval: AI* model is a set of features presented in tabular form on the left side of Fig. 1. Each row represents a person as vector, consisting of 34 features. The model's outcome is a probability of accepting or rejecting the credit, visualized on the right side of the figure. The problem with this approach is the lack of transparency in the decision-making process. The model makes predictions based on specific features without revealing the underlying decision process for the individual case. Providing an explanation of the decision-making process can increase the transparency of the prediction model. For instance, an individual denied credit could ask, "Why was the credit rejected?". Additionally, there may be laws and regulations that require accountability for automated decisions. XAI algorithms can help bridge the transparency gap by providing optimal support for people in decision-making positions, such as credit approval. The research questions (RQx) as well as the hypothesis (Hx) of this paper are as follows.

- RQ1: How similar are the explanations provided by XAI approaches for inputs that differ by no more than 1% in one feature?
- H1: If the inputs are almost identical, then it is expected that only minor changes appear in the explanation. This can be used as a scoring metric for the stability of a XAI algorithm.
- RQ2: Is there a correlation between scaling a selected feature by a factor of a and the resulting feature importance value?
- H2: Scaling a feature by a certain degree, results in a correlation with the feature importance movement.

3 Background and Related Work

Minh *et al.* [17, p. 3511] define XAI in their survey paper as "the study of explainability and transparency for socio-technical systems, including AI." There are several taxonomies according to [4,7,15,21] to classify XAI methods. The focus of this work lies on the evaluation of model-agnostic and model-specific XAI approaches.

3.1 LIME by Ribeiro et al. 2016 [20]

The idea behind LIME is to consider the local model as a black box model. The mode of operation of LIME is based on perturbing an original data point as input into the black box model and using the resulting predictions to train an interpretable surrogate model, which locally approximates the predictions of the black box model. The explanation provided by LIME is defined by:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{1}$$

In Eq. 1, ξ is the explanation of instance x, which is obtained through an optimization task. Function g is an interpretable local model and G is a class of potentially interpretable models. The function f is the original predictor, and π_x defines the radius of the neighborhood around instance x. $\mathcal{L}$ is the loss function that measures the accuracy of the prediction from instance x with respect to the interpretable model g and the original prediction by f in the area of π_x around the original prediction. Ω is a complexity measure of g and serves as a penalty function. LIME calculates feature importance values that show the contribution of each feature for and against a prediction in a certain classification category.

3.2 SHAP by Lundberg and Lee 2017 [16]

The idea of the SHAP has its origins in the game theory. It calculates the extent to which a coalition (set of features) contributes or does not contribute to a particular classification based on the so-called Shapley values. The used implementation was the framework by Lundberg and Lee [16], known as SHap Additive exPlanations. The following definition describes the generation of explanations by the algorithm [16,18]:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{2}$$

The function g describes the explanatory model and z' describes the data instance to be interpreted. Here, z' may have only a subset of all features. The explanation is generated by a linear model where $\phi_i \in \mathbb{R}$ and z'_i is either zero or one for representing the presence or absence of a value from the feature set

z' [16]. Lundberg and Lee provide in their framework (implementation) different explainer models. The computational model approximates Shapley values with perturbations of z'. By that, the complexity problem of computing Shapley values could be solved. SHAP also generates feature importance values like LIME and the expressive power are the same.

3.3 CIU by Främling Initially 1996 [9]

The third model-agnostic algorithm is based on Decision Theory, more specifically on the sub-domain of Multiple Criteria Decision Making (MCDM). In contrast to LIME and SHAP, this approach distinguishes between the measured importance and the utility of an attribute. On the premise of the feature relevance, the focus lies on the contextual importance. This is described in [9,11]:

$$CI_j(\vec{C}, \{i\}) = \frac{C_{max_j}(\vec{C}, \{i\}) - C_{min_j}(\vec{C}, \{i\})}{absmax_j - absmin_j} \tag{3}$$

The explanation model CIU calculates with the function CI_j the importance of feature i in the feature vector $\vec{C}$ for an output label (value) j. The function C_{max_j} determines the maximum output of the prediction j for a certain feature i. The C_{min_j} calculation follows a similar approach. The functions $absmax_j$ and $absmin_j$ determine the highest and lowest prediction from the given data set. More details are provided in the paper by Kary Främling [9–11].

3.4 IG by Sundararajan et al. 2017 [22]

Integrated Gradients is a model-specific approach that differs from the other XAI methods and serves as a verification method. Basic axioms are defined that the XAI approach must fulfil. The most important of these are "Sensitivity", "Implementation Invariance" and "Completeness". The details can be found in the paper, the following describes the parts of the function [22]:

$$\text{IntegradGrads}_i(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha \tag{4}$$

In Eq. 4, F represents the predictor model, x is the input instance, and x' is the baseline input instance, which stands for a neutral prediction, such as a black image. The feature of the input instance is selected by choosing the dimension (feature) of i. $\frac{\partial F(x)}{\partial x_i}$ is the gradient along the predictor of the i-th dimension. Integrated Gradients takes into account the difference between a feature of the input instance and the baseline instance, which is then multiplied by the integrated gradients value of that particular feature.

4 Methodology

According to Bruckert et al. [5], the combination of global and local XAI algorithms is key to ensure the confidence and performance of future ML models.

Building upon this idea, a framework is proposed that utilizes a combination of model-agnostic methods, contrasting and comparing it with a model-specific approach. The main goal is hereby to further analyse the quality of certain XAI explanations by conducting tests with very similar data samples.

Considering the complexity in evaluating a unsupervised model, an approach was developed to get quality insights by conducting changes in the test dataset. The underlying assumption is hereby that minimal changes in the test data, should result in similar explanations. That means, several data sample pairs are generated, which differ in one feature by around 1%. The XAI algorithm is then measured based on the similarity of the resulting explanations. A high variation leads to poor results and less trust. On the other hand, the smaller the variation of the feature importance value (FIV) among the data sample pairs are, the better and more trustworthy is the XAI algorithm.

$$\text{Similarity} = \frac{\text{FIV}_{\text{original}} \cdot 100\%}{\text{FIV}_{\text{perturbated}}} \tag{5}$$

The Eq. 5 shows the formula used to compute the degree of similarity among the data sample pairs. A 100% similarity means hereby that the explanations are identical and that the feature change had no impact on the explanation. In contrast, a value bigger or smaller 100% means that the importance value for a certain feature changed. Besides the similarity computation, the Spearman rank correlation coefficient was applied for the results in RQ2 due to the sake of illustration. The value ranges hereby from -1 to 1 and represents the monotonic relationship of the data. A value of 1 means that by increasing the first part of the data, the second part increases too. In contrast, -1 indicates that while increasing the first part of the data, the second one decreases.

The credit approval task was chosen in the context of AI for critical domains. The "Credit Approval Data Set" from [8] was used for training and testing, but the original dataset only contained acronyms in the column headers for privacy reasons, making it difficult to interpret. To address this, we obtained a cleaned version of the data from Kaggle [6] and used the column headers provided there. For the predictor, a simple neural network was implemented using TensorFlow with two layers, each consisting of 32 neurons, and trained using the Adam optimizer. The output of the model was generated using a softmax function. To interpret the prediction using IG, the same network architecture was used and trained with a sigmoid function in the output layer.

For the use case of the credit approval system that requires explainable AI, a suitable XAI approach must be selected. The approach should be easy to integrate into the existing system (Req. 1), provide explanations of the predictor's behavior (Req. 2), and allow examination of individual features (Req. 3). Post-Hoc models fulfill the first requirement since they do not require changes to the predictor or data. They produce Feature Importance Values for explaining the impact of single features or interdependence between features, meeting the Req. 2.

Table 1. Comparison of the requirements for XAI methods.

Method	LIME	SHAP	CIU	IG
Requirement 1–3	Fulfills	Fulfills	Fulfills	Fulfills
Model Access	No	No	No	Yes
Scope	Local	Local/Global	Local	Local
Approach	Functional	Game Theory	Decision Theory	Functional

As shown in Table 1, each introduced algorithm could be used for evaluating the FIV. The only distinction among the algorithms is whether the predictor is accessible or not. In most cases where a service is used, the predictor may not be accessible, and so LIME, SHAP, and CIU could be chosen. If access to the ML model is granted, then IG is also an option.

The reference implementations of LIME and SHAP from [20] and [16] were used respectively. The CIU implementation presented in [2] and the implementation from the Alibi collection [14] were used as the reference implementations of CIU and Integrated Gradients. For the experiments, the standard arguments of the XAI algorithms, which are listed in Table 2, were used. The library versions used in the experiments are also listed.

Table 2. Overview of parameters for XAI approaches.

Used explainer (data specific)	Input data (type of data)	Model (required)	Optional parameters	Lib. version
LimeTabular Explainer (Yes)	data point, (tabular)	Yes	Top labels: 1 Samples: 5000 Features: 34	0.2.0.1
SHAP Explainer (No)	data point, (tabular)	Yes	–	0.41.0
determine_ciu (No)	data point, (tabular)	Yes	List of min. and max. values	0.0.3
IntegratedGradients (No)	data point, (tabular)	Yes	n_steps: 50 method: gausslegendre baselines: None	0.9.1

Since the credit approval dataset contains nominal variables (such as "Industry" and "Ethnicity"), a one-hot encoding was applied, resulting in a dataset with 690 samples and 34 features each. To train and test the deep neural network, the data was splitted (80/20) and transformed along the columns using the StandardScaler from scikit-learn, resulting in an average of 0 and a standard deviation of 1. The same preprocessing was applied to the test data used for the XAI algorithms. RQ1: To analyze the similarity of explanations for almost identical data sample pairs, 100 data samples were selected and further processed. Four

copies of each sample were created, with each copy having one column changed by +1%. The columns "Age", "Debt", "YearsEmployed", and "Income" were chosen for this purpose, as they contain continuous numerical data. The data of these features is in decimal form and therefore possible to change by around 1%. After the feature changes, the data was merged into a dataset with a total of 500 rows and 34 columns. RQ2: In contrast to RQ1, RQ2 conducts changes on just one column and 10 data samples. These have six copies each, representing a feature change of +1%, +5%, +10%, −1%, −5%, and −10%. The objective is to analyze the relationship between the feature change and the similarity changes of the explanations. The merged dataset has 70 rows, where rows 10–70 represent the changed data samples. As only one feature-column is changed, the other 33 remain the same throughout the copies. The "Age" and "Income" columns are changed, resulting in two datasets, each with a shape of (70,34). After obtaining the explanations from the XAI models, the data was shifted on a positive scale. To avoid any problems for the following similarity calculation of the data pairs, a minimum value of $1 \cdot 10^{-12}$ was chosen. A data pair refers to a data sample that was created through minimal changes in one column. For RQ1, we have 100 original data samples, each of which has four pairs (total 400 pairs). In contrast, the dataset in RQ2 originated based on 10 data samples, each of which has six pairs (total 60 pairs/dataset).

5 Results

As mentioned in Sect. 4, the explanation similarity analysis for RQ1 is based on the feature changes in the column "Age", "Debt", "YearsEmployed" as well as "Income". The result is hereby a similarity matrix of 400×34, which represents the degree of similarity for all data pairs and their feature values.

Due to the size of the result set (for each algorithm 400×34), the explanation similarities of the "Income" column are presented in Fig. 2. The figure illustrates the distribution of the similarities for each XAI algorithm, where it can be clearly seen that most of the explanation pairs have a similarity of 99 to 101%. That means, a feature change of around 1% resulted in a explanation similarity of mostly 99 to 101% for the "Income" column. By comparing the results of the algorithms which each other, it can be said that especially SHAP has great results.

The same applies for CIU, which performs slightly worse then SHAP. Most of the samples lie in a similarity range of 99–101%, while the amount of samples lower or higher then that is very small. Even though Integrated Gradients achieved comparable results, a higher data dispersion as well as a few outliers could be observed. In contrast to the results of CIU, SHAP as well as Integrated Gradients, the results of LIME are highly distributed and less centered. As illustrated in Fig. 2, small changes in the features can lead to big differences in terms of the explanations. According to the quality approach mentioned in Sect. 4, this makes it unreliable and less trustworthy. On the other side, CIU, Integrated Gradients and especially SHAP deliver good results.

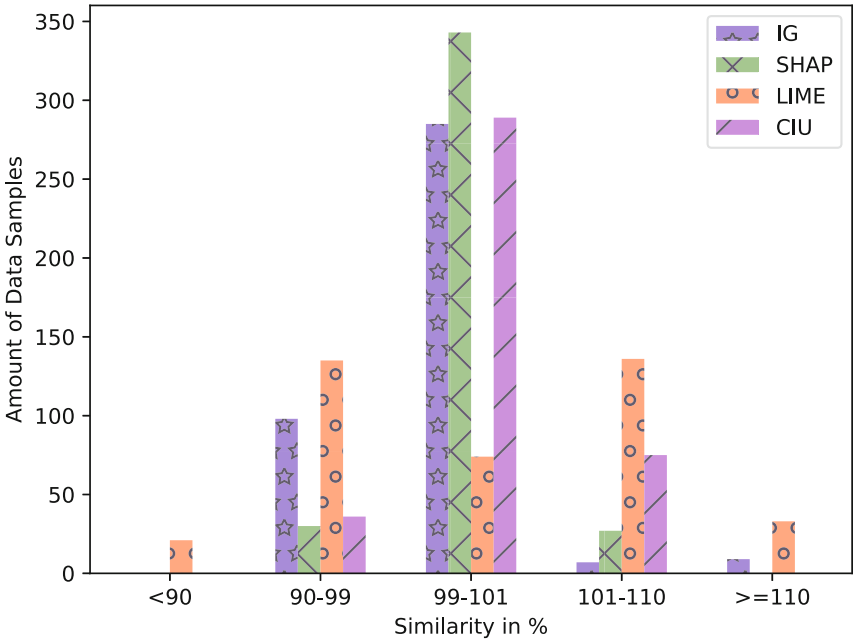

Fig. 2. Algorithm Comparison about the similarity of the explanations for the feature "Income" - representation of the feature importance similarities for the data pairs. The similarity measures the percentage match between the explanation of the original data sample as well as the one with a feature change of around 1%. The results of "Integrated Gradients", "SHAP", "LIME" and "CIU" are compared.

However, these conclusions relate to the column "Income" and can change among the different features. For instance in Fig. 3, it can be seen that the explanation similarity analysis for the column "Age" leads to slightly different results. The dispersion of the data got larger, especially for the results of LIME and IG.

In contrast to RQ1, RQ2 focuses on a relation between an increasing feature change as well as the similarity in the explanations. That means, if increasing or decreasing the value of a feature, has the same impact on the feature importance values. Table 3 compares hereby the similarity of the explanations. If the similarity is above 100%, it means the feature importance shrinked and vice versa. Except for the results of CIU, no clear relation of the changes in the feature values and their resulting feature importance values could be observed. For the values of CIU however, a monotonic relationship could be analysed. That means for instance, the change of the "Income" feature values caused a similar feature importance movement in "Age" for the CIU algorithm. By considering that every decision is based on 34 features, the change of the feature importance value can be considered as big. This is crucial by comparing the results with each other,

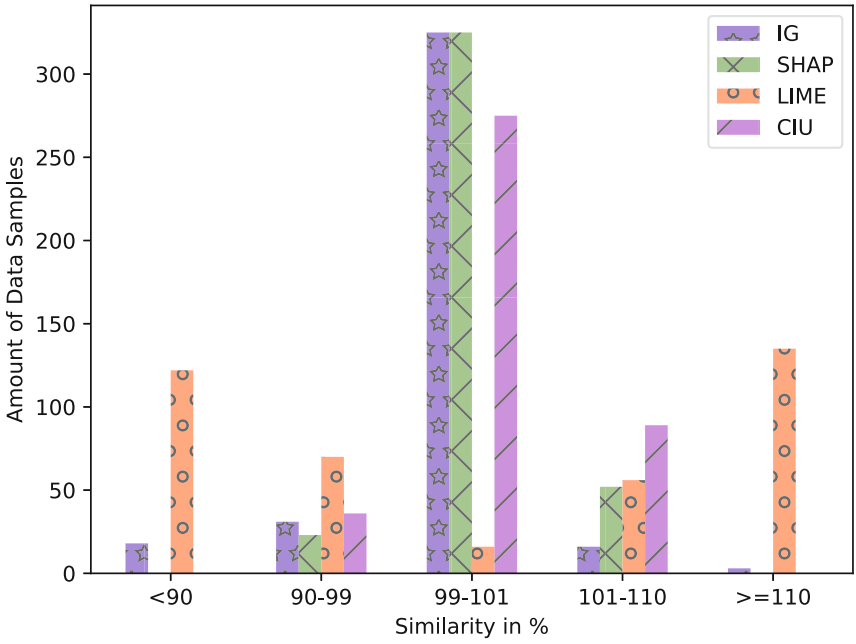

Fig. 3. Algorithm Comparison about the similarity of the explanations for the feature "Age" - representation of the feature importance similarities for the data pairs. The similarity measures the percentage match between the explanation of the original data sample as well as the one with a feature change of around 1%. The results of "Integrated Gradients", "SHAP", "LIME" and "CIU" are compared.

Table 3. Feature Changes for "Income" - Analysis of the relation between the varying values of feature "Income" and the changing explanation similarities of the feature "Age".

	Feature Change	Changes in Income Column					
		+1%	+5%	+10%	−1%	−5%	−10%
Age	LIME	81,87	93,60	95,25	193,12	157,12	105,29
	SHAP	99,98	100,05	100,02	99,96	99,82	99,87
	CIU	100,70	102,23	105,38	99,64	97,79	95,83
	Integrated Gradients	94,93	93,74	92,50	102,56	99,04	99,06

since a lot of times when similar movements occur, the similarity changes are not significant.

This can be seen in Table 4, which further processes the data by computing the Spearman rank correlation coefficient. Details about the exact feature importance changes are missing in Table 4 and therefore make it hard to distinguish similar values. As mentioned in Sect. 4, the values range from -1 to 1 and represent the monotonic relationship. In comparison to LIME, SHAP and

Integrated Gradients, CIU shows a clear monotonic relationship in all of the columns. Besides positive monotonic relationships, CIU also has some negative monotonic relationships. That means, a change in the feature value can either result in feature importance values that change in the same (positive) or opposite (negative) direction. Besides the results of CIU, LIME and Integrated Gradients had a monotonic relationship with some of the features.

Table 4. Feature Changes for "Age" and "Income" - analysis of the relation between varying feature values and their changing explanation similarities. Results are presented with the computed Spearman rank correlation coefficient.

	Feature Change	Changes in Income Col.		Changes in Age Col.	
		+1%, +5%, + 10	−1%, −5%, −10%	+1%, +5%, +10%	−1%, −5%, −10%
Age	LIME	1.0	1.0	−0.5	−1.0
	SHAP	0.5	0.5	−0.5	−1.0
	CIU	1.0	1.0	1.0	1.0
	Integrated Gradients	−1.0	0.5	−1.0	−0.5

6 Conclusion

The first aspect of the research was to analyse the similarity of explanations for slightly different test instances (RQ1), which can be understood as an metric approach to evaluate the trustworthiness of certain XAI algorithms. For the investigated algorithms (CIU, Integrated Gradients, LIME, and SHAP), it was hypothesized that minimal feature deviations would only lead to very small changes in the feature importance values. SHAP performed consistently well due to its theoretically well-founded calculation model, which tries many coalitions to calculate an objective interpretation. Molnar et al. also considers SHAP to be highly robust in terms of legal traceability [18]. Due to it's model-specific procedure, which uses both the I/O behavior of the predictor and the internal gradients to generate explanations, Integrated Gradients could achieve great results. However, in contrast to SHAP and CIU, a larger dispersion of the data as well as more outliers could be observed. The results of LIME were highly distributed and therefore the least trustworthy model. One possible reason for this is the local fidelity of LIME and the random generation of neighbor instances used to train the interpretable proxy model. Based on the similarity metric, CIU, IG, and SHAP consistently rank high, but also show some noticeable deviations. Focusing solely on the "High Accuracy Similarity Range" (i.e., 99% to 101%) is not enough to provide a definitive answer. It is crucial to take into account the outliers and the quantity of non-centered data. This could potentially form the foundation for future research. The range and number of outliers with respect to similarity give a more objective overview. This approach could also be utilized

as a starting point for creating an enhanced metric that objectively assesses the efficacy of XAI algorithms.

In RQ2, the aim was to examine the correlation between changing feature importances and the changed feature values. CIU showed hereby a clear correlation. On the other hand, there was only a partial correlation found for LIME, IG, and SHAP. Therefore, the initial hypothesis can not be confirmed. One possible explanation for this could be the interdependencies between the features, which were not investigated.

Finally, the evaluation of the ML model is beneficial in critical areas such as finance. The evaluation of these black box predictive models can reveal the internal processes by explaining the decision path. It has been shown that explanatory approaches such as CIU, Integrated Gradients and SHAP produce more precise explanations, according to the presented metric, than LIME.

References

1. Amro, B., Najjar, A., Macido, M.: An intelligent decision support system for recruitment: resumes screening and applicants ranking (2022)
2. Anjomshoae, S., Kampik, T., Främling, K.: Py-CIU: a python library for explaining machine learning predictions using contextual importance and utility. In: IJCAI-PRICAI 2020 Workshop on Explainable Artificial Intelligence (XAI) (2020)
3. Anshari, M., Almunawar, M.N., Masri, M., Hrdy, M.: Financial technology with AI-enabled and ethical challenges. Society **58**(3), 189–195 (2021). https://doi.org/10.1007/s12115-021-00592-w
4. Arya, V., et al.: One explanation does not fit all: a toolkit and taxonomy of AI explainability techniques. CoRR abs/1909.03012 (2019). http://arxiv.org/abs/1909.03012
5. Bruckert, S., Finzel, B., Schmid, U.: The next generation of medical decision support: a roadmap toward transparent expert companions. Front. Artif. Intell. **3**, 507973 (2020)
6. Cortinhas, S.: Credit card approvals (clean data) from kaggle (2022). https://www.kaggle.com/datasets/samuelcortinhas/credit-card-approval-clean-data. Accessed 16 Apr 2023
7. Das, A., Rad, P.: Opportunities and challenges in explainable artificial intelligence (XAI): a survey (2020). https://doi.org/10.48550/ARXIV.2006.11371. https://arxiv.org/abs/2006.11371
8. Dua, D., Graff, C.: UCI machine learning repository (2017). http://archive.ics.uci.edu/ml. Accessed 16 Apr 2023
9. Främling, K.: Decision theory meets explainable AI. In: Calvaresi, D., Najjar, A., Winikoff, M., Främling, K. (eds.) EXTRAAMAS 2020. LNCS (LNAI), vol. 12175, pp. 57–74. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-51924-7_4
10. Främling, K.: Explainable AI without interpretable model. CoRR abs/2009.13996 (2020). https://arxiv.org/abs/2009.13996
11. Främling, K.: Contextual importance and utility: a theoretical foundation. In: Long, G., Yu, X., Wang, S. (eds.) AI 2022. LNCS (LNAI), vol. 13151, pp. 117–128. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-97546-3_10

12. Guo, H., Polak, P.: Artificial intelligence and financial technology FinTech: how AI is being used under the pandemic in 2020. In: Hamdan, A., Hassanien, A.E., Razzaque, A., Alareeni, B. (eds.) The Fourth Industrial Revolution: Implementation of Artificial Intelligence for Growing Business Success. SCI, vol. 935, pp. 169–186. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-62796-6_9
13. Kaur, C., Garg, U.: Artificial intelligence techniques for cancer detection in medical image processing: a review. Mater. Today Proc. **81**, 806–809 (2021)
14. Klaise, J., Looveren, A.V., Vacanti, G., Coca, A.: Alibi explain: algorithms for explaining machine learning models. J. Mach. Learn. Res. **22**(181), 1–7 (2021). http://jmlr.org/papers/v22/21-0017.html
15. Liao, Q.V., Singh, M., Zhang, Y., Bellamy, R.: Introduction to explainable AI. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–3 (2021)
16. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS 2017, pp. 4768–4777. Curran Associates Inc., Red Hook (2017)
17. Minh, D., Wang, H.X., Li, Y.F., Nguyen, T.N.: Explainable artificial intelligence: a comprehensive review. Artif. Intell. Rev. **55**, 3503–3568 (2021)
18. Molnar, C.: Interpretable machine learning (2022). https://christophm.github.io/interpretable-ml-book/. Accessed 16 Apr 2023
19. Reddy, S., Allan, S., Coghlan, S., Cooper, P.: A governance model for the application of AI in health care. J. Am. Med. Inform. Assoc. **27**(3), 491–497 (2020)
20. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should i trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2016, pp. 1135–1144. Association for Computing Machinery, New York (2016). https://doi.org/10.1145/2939672.2939778
21. Speith, T.: A review of taxonomies of explainable artificial intelligence (XAI) methods. In: 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 2239–2250 (2022)
22. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, vol. 70, pp. 3319–3328. JMLR.org (2017)