applied system innovation

MDPI

Article

Evaluation Metrics Research for Explainable Artificial Intelligence Global Methods Using Synthetic Data

Alexandr Oblizanov [1], Natalya Shevskaya [1], Anatoliy Kazak [2,*], Marina Rudenko [3] and Anna Dorofeeva [2]

[1] Faculty of Computer Science and Technology, Saint Petersburg Electrotechnical University Leti, 197376 Saint-Petersburg, Russia
[2] Humanitarian Pedagogical Academy, V.I. Vernadsky Crimean Federal University, 295007 Simferopol, Russia
[3] Institute of Physics and Technology, V.I., Vernadsky Crimean Federal University, 295007 Simferopol, Russia
* Correspondence: kazak@cfuv.ru or kazak_a@mail.ru

Abstract: In recent years, artificial intelligence technologies have been developing more and more rapidly, and a lot of research is aimed at solving the problem of explainable artificial intelligence. Various XAI methods are being developed to allow the user to understand the logic of how machine learning models work, and in order to compare the methods, it is necessary to evaluate them. The paper analyzes various approaches to the evaluation of XAI methods, defines the requirements for the evaluation system and suggests metrics to determine the various technical characteristics of the methods. A study was conducted, using these metrics, which determined the degradation in the explanation quality of the SHAP and LIME methods with increasing correlation in the input data. Recommendations are also given for further research in the field of practical implementation of metrics, expanding the scope of their use.

Keywords: explainable artificial intelligence; XAI; explanation metrics; synthetic data

Citation: Oblizanov, A.; Shevskaya, N.; Kazak, A.; Rudenko, M.; Dorofeeva, A. Evaluation Metrics Research for Explainable Artificial Intelligence Global Methods Using Synthetic Data. *Appl. Syst. Innov.* **2023**, *6*, 26. https://doi.org/10.3390/asi6010026

Academic Editor: Friedhelm Schwenker

Received: 3 January 2023
Revised: 16 January 2023
Accepted: 5 February 2023
Published: 9 February 2023

1. Introduction

One of the areas for the development of AI (Artificial Intelligence) technologies is the development of explainable artificial intelligence (hereinafter referred to as XAI) methods that allow users to understand why machine learning algorithms have come to certain results and conclusions. These methods are mainly aimed at increasing user confidence in AI technologies, but their imperfection undermines this trust [1]. For example, many XAI methods (LIME, GGCAM, OCC, etc.) can incorrectly interpret the model if the input data is distorted: for example, if the colors of a small number of pixels in several patterns are changed in the input data of an image classification model [2].

For the practical application of XAI methods, it is important to have an idea about their speed, accuracy, and resource intensity. When evaluating XAI methods, different problem-oriented approaches are used [3–5], which often do not have the portability property [6]. The evaluation is aimed at certain parameters of the method, ignoring the rest, and this becomes a confirmation bias [7]. To solve these problems, it is possible to define a set of metrics that take into account a wide range of technical characteristics of the XAI method (for example, computational complexity, accuracy [8]).

XAI methods are divided into local (providing an explanation for only a part of the units in the data set) and global (providing an explanation for the entire data set). In this paper, we consider the problem of evaluation global XAI methods. As models to which the method will be applied, models of artificial neural networks will be used. Among the global XAI methods, several of the most famous were chosen as the object of study—SHAP, LIME.

The goal of the work is to develop metrics that allow evaluating the result of applying the methods of explainable artificial intelligence LIME, SHAP to machine learning models using synthetic data.

To achieve the goal, must be completed the following tasks:

1. Determination of criteria for approaches to assessing XAI methods;
2. Study of existing metrics for evaluating XAI methods;
3. Selection of XAI evaluation metrics and their modification;
4. Development of software for generating synthetic data and calculating metrics;
5. Study of SHAP, LIME XAI methods using the developed software

The object of the research is the methods of explainable artificial intelligence. The subject of the study is the metrics for evaluating the methods of explainable artificial intelligence.

2. Relevance. Overview of Analogues

To search for analogues of solving the problem of estimating global methods for explaining machine learning models, we searched articles in Google Scholar for the following keywords and phrases: (1) Explainable Artificial Intelligence; (2) Interpreting model predictions; (3) XAI benchmark; (4) XAI evaluation; (5) Metrics for XAI; (6) Method evaluation. To select relevant articles, a filter was used by the year of issue: 2019–2021.

2.1. Determination of Criteria for Evaluating Analogues

In order to compare analogues, it is necessary to understand what key criteria can characterize the system for evaluating XAI methods. The criteria should mainly be based on the metrics used in the system. In this case, a metric is a measure that allows one to obtain the numerical value of some property of the algorithm, its implementation and practical application. To evaluate the methods of explainable artificial intelligence, a variety of metrics from various sciences can be used: mathematics, computer science, sociology, etc.

Some methods suggest specifying method requirements for evaluation. In this case, the requirements should be such that it is possible to define a metric showing the extent to which the assessment system complies with this requirement.

An important criterion is the number of metrics that have a methodical or mathematical description. If an assessment method offers metrics that do not have a precise description of how they are calculated, researchers may calculate these metrics in different ways, which will not solve the problem of subjectivity and the inability to compare different assessments of XAI methods. Under the methodological description, which is most often applicable to sociological metrics, is meant a step-by-step instruction on measuring a certain metric, indicating the input parameters, the method for obtaining them and forming the metric for them. For example, a methodological description of the measurement of a metric could be conducting interviews and questionnaires using the Likert scale with questions from the checklist.

The number of metrics that depend on the size of the sample is a criterion that allows you to assess the complexity of the implementation of the assessment method and its dependence on the size of the study. A sample is a set of sets of input parameters required to obtain a set of metrics. This criterion is based precisely on the description of the methods—whether there are requirements for the sample size, whether the dependence of the quality of the metric on it is mentioned.

The evaluation metrics of XAI methods can be roughly divided into two groups:

1. Metrics based on technical specifications
2. Metrics based on sociological and cognitive characteristics

The technical characteristics of the XAI method are its measurable parameters and properties that determine the software implementation, applicability on a computer, the mathematical properties of the algorithms used in it and the method as a whole. A metric that characterizes one or more technical characteristics allows you to evaluate the method from the point of view of constructing and analyzing algorithms, mathematical analysis. For example, such a metric can be the computational complexity of the method, or the dependence of the number of operations on the computational complexity of the model to which the XAI method is applied; accuracy of parameter determination; limitations and vulnerabilities. To determine the number of such metrics, a criterion is defined—the number of evaluated technical characteristics. The performance of explanation methods

depends on many factors: the machine learning model they explain and the data on which the model was trained. Accounting for heterogeneity of input data criterion shows if the methods of evaluation explore the performance differences of the method under different models and datasets.

Sociological and cognitive characteristics determine the user's interaction with the XAI method, the receipt and assimilation of information provided by the methods, if generalized, the degree and nature of the influence of the method on the user. However, such metrics, as it is right, significantly depend on the sociological sample: age, computer proficiency, degree of familiarity with AI technologies, etc. Thus, the metrics depend significantly on the scope of the study. To take this into account, a criterion of the amount of research required to calculate the metrics is defined.

There are several ways to calculate the metrics described in Table 1. The list of criteria is presented in Table 2.

Table 1. Methods for calculating metrics.

Method Name	Characteristic
Functionality-grounded	Theoretical, based on the description of the algorithm
Human-grounded	Experimental calculation
Application-grounded	User Experience Research

Table 2. Criteria for comparing analogues.

No	Criterion Name
1	The number of metrics that have a mathematical or methodological description
2	Number of evaluated technical characteristics
3	Accounting for heterogeneity of input data
4	Types of studies for calculating metrics
5	The amount of research required to calculate the metrics

2.2. Description of Analogues

A study by Keane M. T. et al. [9] raises the issue of evaluating counterfactual explanatory methods based on psychological issues and seeks evidence that the nature of evaluations of certain explanatory methods correlates with the user experience of using these methods. The article also points out that such an analysis is relevant not only for counterfactual methods of explanation, since many other types of XAI are not able to properly meet the requirements of end users. To evaluate the methods of explanation, the authors propose certain requirements for the method:

1. The method must be accuracy-oriented (proximity-guided)
2. The method should focus on functionality (eng. focused on features)
3. The method must be distributionally stable (eng. distributionally-faithful)
4. The method must be instance-guided

The proposed four requirements make it possible to assess the technical characteristics of the method. It can also be noted that these requirements are quite relevant for users. However, any methodological description of how the compliance of the method with the requirements should be determined is not indicated in the article, and user experience is not provided.

The article by Rosenfeld A. [7] considers the problem of evaluating the methods of explainable artificial intelligence and sets the goal of showing how the ambiguity in determining the goals that XAI is designed to achieve affect the qualitative and quantitative evaluation of these methods. The article mentions that modern XAI methods are often developed with the aim of maximizing their performance, including using artificial intelligence methods, even if this may affect the stability of the explain method algorithm.

The article proposes a set of metrics that can be used to evaluate XAI methods:

1. D, performance difference between models and XAI method execution;
2. R, the number of rules in the model explanation (rule based);
3. F, the number of features used to construct the explanation;
4. S, the stability of the explanation of models;

The four metrics proposed by the authors dive deep enough into the technical features of the implementation of the XAI method. They may be useful for further research, however, for example, the R, F metrics for potential users are hardly representative. It should be noted that examples of their mathematical description are given for metrics.

The study by Hsiao J. H. [8] considers the problem of evaluating the quality of explanatory methods and pays special attention to research in the field of cognitive science and psychology, which allow a more pragmatic and naturalistic approach to evaluating XAI methods. The article proposes the following cognitive assessment metrics:

1. Explanation Goodness
2. User satisfaction
3. User Curiosity/Attention Engagement
4. User Trust/Reliance
5. User understanding
6. Productivity/Productivity of use (English User Performance/Productivity)
7. System Controllability/Interactivity

The authors Lin Y. S., Lee W. C., Celik Z. B. [10] conducted a study that addresses the problem of assessing the interpretability of models using XAI methods. It points out that in existing studies, measurements are made by humans and can be tedious and time-consuming, as well as introduce bias and lead to inaccurate estimates. In addition, vulnerabilities were found in XAI methods, which in certain situations could give a similar interpretation of models trained on true and randomized data. The study presents images as input data, and the model to which the XAI method is applied is an artificial neural network. To evaluate the methods, the authors propose to use the masking method. So, the XAI method uses a lot of models trained on various input data distorted according to certain patterns (model trojaning), and the output data of the methods is compared. The paper proposes the following metrics:

1. Recovery Rate—determines the effectiveness of trigger detection
2. Computational Cost—determines the cost of defining triggers
3. Intersection of triggers in relation to their union (eng. Intersection over Union)—also determines the effectiveness of trigger detection
4. Recovering Difference—determines the correctness of the rejection of triggers

In a study by Zhou J. et al. [11] considers the problem of inconsistency of indicators for assessing the quality of explanation of methods, which makes comparisons of XAI methods difficult. Explainability is inherently a very subjective concept that depends on users, and therefore the authors propose a structure of metrics, divided into types and subtypes:

1. Metrics based on method application and user
    a. Subjective metrics
        i. User trust
        ii. User preference
        iii. User confidence
    b. Objective metrics (user psychological signals)
        i. Electrical activity of the skin (eng. Galvanic Skin Response)
        ii. Blood Volume Pulse
2. Metrics based on functionality (for methods for evaluating artificial neural network models)
    a. Number of operations depending on model size
    b. Level of disagreement

c. Interaction strength

The study also points out that different types of metrics should be integrated together. The metrics given in the article are only examples of metrics related to one or another type, while a methodological or mathematical description is not given for them.

2.3. Findings from the Comparison

As a result of the comparison (Table 3), it was revealed that many analogues that offer the evaluation of technical characteristics either do not take into account the heterogeneity of the input data, or do not have a mathematical or methodological description of the calculation of metrics. Nevertheless, the approaches of each of the methods deserve attention and can be supplemented.

Table 3. Comparison of analogues.

Criterion/Analog	The Number of Metrics with Mathematical or Methodological Description	Number of Evaluated Technical Characteristics	Accounting for Heterogeneity of Input Data	Types of Studies for Calculating Metrics	The Amount of Research Required to Calculate the Metrics
№1, Keane M. T. et al., 2021	0	4	Yes	Functionality-, Human-grounded	Middle
№2, Rosenfeld A., 2021	4	4	No	Human-grounded	Middle
№3, Hsiao J. H. et al., 2021	7	0	Yes	Human-grounded	High
№4, Lin Y. S., Lee W. C., Celik Z. B., 2020	4	4	No	Human-grounded	Middle
№5, Zhou J. et al., 2021	0	3	Yes	All	Extremely high

Among the methods based on the technical characteristics of the XAI methods, analogue No. 1 does not offer a methodical or mathematical description of the metrics, instead some requirements for the method are proposed. This approach leads to the fact that when using the evaluation method, the metrics will be calculated in different ways, which does not solve the problem of their incompatibility.

Analogue #4 offers an assessment of the work of methods with synthetic data, in which noise disturbances and artificial trigger features are added. In general, the use of synthetic data allows you to specify the features of the sample—the type and parameters of the distribution of features, the type of function for calculating the true prediction, the addition of noise disturbances or insignificant features. This approach allows us to explore the characteristics of the method depending on the input data, for example, by calculating a certain accuracy metric depending on the correlation of features.

2.4. Solution Method Choosing

Based on the review of analogues and other studies [12–45] and the considered criteria, the main requirements for the solution were compiled:

1. Taking into account the technical characteristics of the XAI method
2. Accounting for heterogeneity of input data using synthetic data
3. Availability of a methodical or mathematical description of each metric
4. Ability to calculate metrics when running methods with different machine learning models

It's important to automate calculation of metrics so it's proposed to develop Metrics Calculation Tool which will support SHAP, LIME methods and will use synthetic data.

Schematic representation of the method is shown in Figure 1.

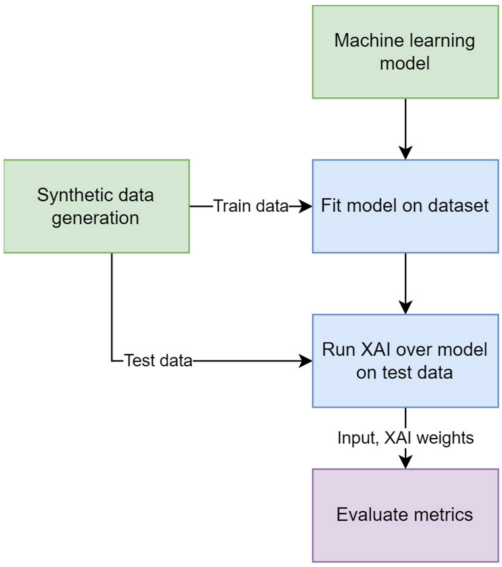

Figure 1. Schematic representation of the method.

3. Experiments

3.1. Metrics Modification

1. Faithfulness metric. The faithfulness metric allows you to determine the degree of correspondence between the explanation of the importance of each individual feature and is determined by the Pearson sample correlation coefficient between the weights of the feature and its approximate contribution to the change in the model prediction when it is removed or fixed.

A schematic representation of the iteration of the algorithm is shown in Figure 2.

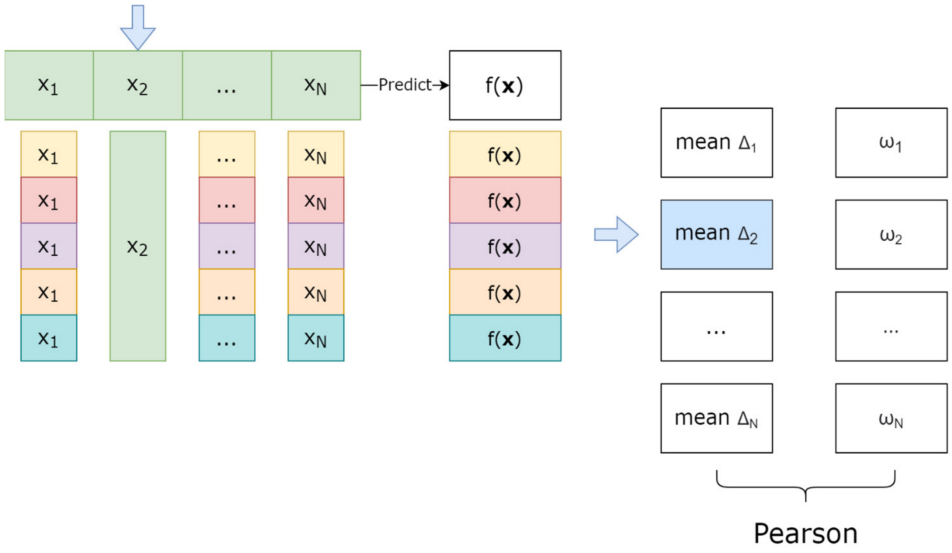

Figure 2. Scheme of the algorithm for calculating the faithfulness metric.

The values of the Pearson correlation coefficient, by definition, are in the range from −1.0 to 1.0, the faithfulness metric can take values from 0 to 1. Values close to 1 indicate the correctness of the distribution of weights, and close to 0 indicate that the influence of features on the prediction does not match the set of weights. It is important to note that this metric does not consider the correctness of each weight, but considers them in aggregate, which in the general case is not always the main indicator of the correctness of an explanation.

2. Monotonicity metric. The monotonicity metric is based on this concept and determines the correctness of a sequence of features, ordered by increasing their weight, obtained by XAI methods. If monotonicity is not observed, then the XAI method allowed a distortion of feature priorities: a feature with less influence received more weight, or vice versa. This metric also cannot assess the accuracy of the weight value of each feature, but it can assess the correctness of the distribution of weights between features.

A schematic representation of the iteration of the algorithm is shown in Figure 3.

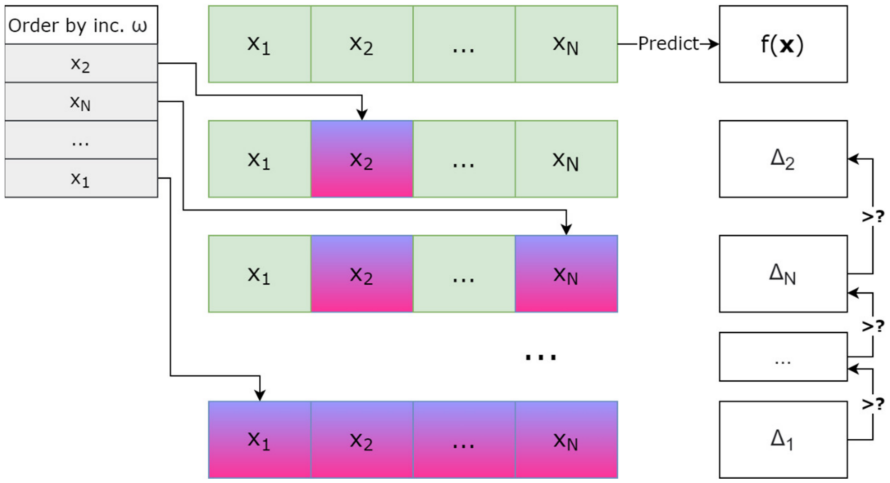

Figure 3. Scheme of the algorithm for calculating the Monotonicity metric.

The main difference between the faithfulness metric and the monotonicity metric is that to calculate the faithfulness metric, subsets of features are considered, from which each of the features is iteratively removed. At the same time, the monotonicity metric considers the cumulative effect of adding features, and therefore can be more reliable if they are highly correlated.

3. Incompleteness metric. The Incompleteness metric determines the effect of noise perturbations of each feature on model predictions by calculating the difference between the weights and the difference between the original prediction and the predictions given by the noisy feature sets.

3.2. Synthetic Data

The generation of synthetic data is divided into two stages: the generation of feature sets and their markers. Feature sets are generated by sampling from a multivariate Gaussian distribution. Noise is added to the feature values by summing their values with a random value from a normal distribution with zero mathematical expectation, the noise factor determines the standard deviation. Markers for feature sets are generated as a result of some feature function. It is possible to use linear, nonlinear and their impurities, piecewise linear functions.

3.3. Implementing the Metrics Calculation Tool

Python version 3.9 was chosen as the programming language. Programming environments used are PyCharm Community and Visual Studio Code. List of used libraries presented in Table 4.

Table 4. Table of used libraries.

Library Name	Description	Library Version
NumPy	A library that implements linear algebra operations, mathematical functions, elements of statistical analysis	1.21.0
Matplotlib	Library for plotting various types of graphs	3.5.1
Scipy	Library designed to perform scientific and engineering calculations	1.8.0
Pandas	Library for working with tabular data structures	1.4.1
Shap	Library with implementation of the XAI SHAP method	0.40.0
Lime	Library with the implementation of the XAI LIME method	0.2.0.1
Scikit-learn	Library with tools for designing and training models	1.0.2

Setting of the designed program launch parameters is carried out using the json files of the experiment configuration and the script. The program has a CLI interface and has an indication of the process of performing experiments. Calculation results are saved in CSV and JSON files. In addition, log files are saved. There is a separate module for plotting dependencies of metric values on parameters and saving them in PNG format, which is launched separately. It is possible to customize the models used, XAI methods, parameters for generating synthetic data, calculated metrics.

There some modules implemented in the program:

1. Synthetic data generation module. The synthetic data generation module includes classes that allow you to generate sets of features by sampling from various distributions (multivariate normal distribution, including conditional distribution according to the method described in paragraph 2.1), as well as markers for data sets using various methods (linear, piecewise -linear, non-linear function) for linear regression or classification models.
2. Module for calculating metrics of XAI methods. Each metric is implemented as a separate class: Faithfulness, Infidelity, Monotonicity. When any of these classes are initialized, an instance of the CustomData class and an instance of the machine learning model are passed. Each of the metric classes implements the evaluate function, which takes as input the initial data set, the weights obtained from the XAI method, and a parameter that controls the number of additionally generated samples for calculating the metric.
3. XAI methods application module. The module is responsible for initializing and applying the XAI methods. For the SHAP method, two classes are implemented that provide an interface for calling the shap.Explainer and shap.KernelExplainer classes included in the shap library. The library is the official implementation of the method, and its documentation contains many examples of applying the method to models of various types.

3.4. Results and Discussion

Based on the results of the experiment, it is also possible to build graphs of dependence on other parameters, for example, on the sample size, or plotting the stability of metric values, all other things being equal. Dependence is built by a piecewise line graph, and the area from the minimum to the maximum value of the metric is also painted over.

As a result of the experiment with a given configuration, the values of the metrics were obtained. For example, the values of the faithfulness metric for the linear regression model are presented in Table 5.

Table 5. Confidence metric for a linear regression model.

Method	Marker Function	The Value of the Metric at the Correlation Coefficient					
		0.0	**0.2**	**0.4**	**0.6**	**0.8**	**0.99**
SHAP	Linear	0.94	0.86	0.77	0.66	0.66	0.6
	Non-linear	0.96	0.87	0.81	0.81	0.76	0.62
LIME	Linear	0.86	0.77	0.72	0.65	0.66	0.63
	Non-linear	0.92	0.88	0.8	0.8	0.72	0.57

Graphs of dependences of the value of the faithfulness metric on the correlation coefficient are shown in Figure 4.

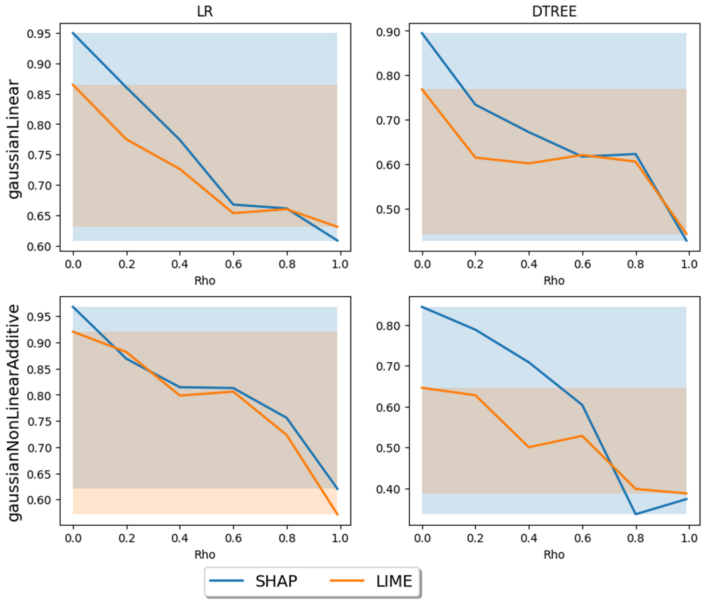

Figure 4. Dependence of the reliability metric on the correlation coefficient.

With correlation coefficients of 0.8 and especially 0.99, both methods show fairly significant errors. It is worth noting the rather high values of the metric for the non-linear marker generation function when applying methods to a linear regression model—methods are ways to identify non-linear dependencies and evaluate their contribution to model prediction.

Graphs of dependences of the value of the monotonicity metric on the correlation coefficient are shown in Figure 5.

It's worth noting that the XAI methods also need tweaking. For example, the stability and accuracy of the explanation of the Lime method can vary significantly with different kernel settings. The SHAP method has many modifications that differ in the way the Shapley values are calculated and taken into account. Additional studies are needed to study the effect of various parameters and method modifications.

Depending on the other metrics, the incompleteness metric determines the correspondence of weights to prediction changes when noise is added to all features at once. The value of the metric itself is the average difference between these values, so the smaller the metric, the more complete the explanation can be considered. It can be seen from the graphs that with an increase in the correlation coefficient, the explanation becomes more complete. Higher values of the metric for binary tree models are also noticeable. In general, the values of the metric are low, which indicates the correct determination of the total contribution of all features to the model prediction.

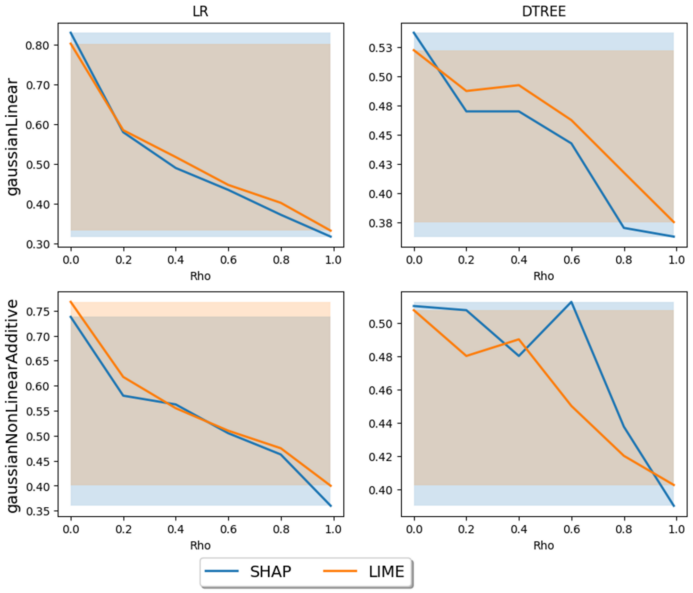

Figure 5. The dependence of the monotonicity metric on the correlation coefficient.

The results of calculating the incompleteness metric are shown in Figure 6.

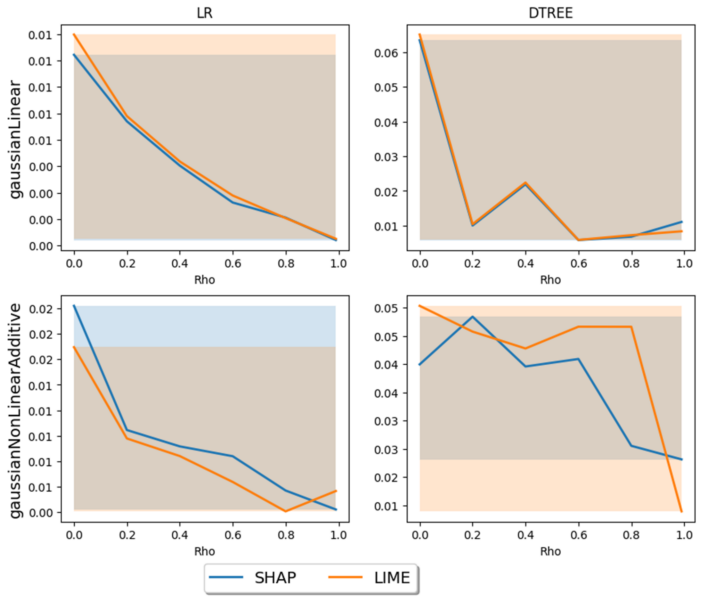

Figure 6. The incompleteness metric.

4. Conclusions

As a result of the study, the criteria for the evaluation system of XAI methods were determined. For a comprehensive assessment, it is necessary to take into account both the technical characteristics of the method, which can be determined experimentally or mathematically, and the sociological and cognitive characteristics that can be obtained in the course of sociological research and are strongly influenced by sampling parameters. At the same time, a balance should be observed between the types of metrics in order to eliminate the bias of the assessment.

The considered analogues [7–11] with methodological and mathematical descriptions of metrics assumed the use of either technical or cognitive characteristics. In analogue [11], the importance of their joint use in the assessment was mentioned, but specific metrics were not indicated.

From the study, the following conclusions can be drawn:

- SHAP and LIME methods are comparable in terms of accuracy of explanation.A study was conducted, using these metrics, which determined the degradation in the explanation quality of the SHAP and LIME methods with increasing correlation in the input data;
- The completeness of explanations of both methods is satisfactory, but the weights of each feature separately, as well as the order of increasing weights, largely depend on the input data and the machine learning model;
- Both methods provide less accurate explanations as the correlation coefficient of features in the input data increases, with a correlation of >0.5, the explanations of the methods are unstable and may be questioned;
- Both methods provide less accurate explanations when applied to decision tree models than when applied to linear regression models;
- The SHAP method shows comparable explanatory accuracy with a linear and non-linear marker function, while the LIME method is less accurate with a non-linear function;
- The execution time of the LIME method is more than 10 times the execution time of the SHAP method.

Further research on the basis of the developed tool for calculating metrics for evaluating the methods of explainable artificial intelligence can be aimed at:

- Expansion of the list of XAI methods, for example, methods MAPLE, ER-SHAP, Breakdown;
- Study of evaluation metrics of XAI methods when applied to classification models;
- Study of dependence of SHAP method evaluation metrics on kernel parameters;
- Expansion of the list of machine learning models, for example, models of autoencoder, deep neural network, vector machines, convolutional neural network.

Author Contributions: Conceptualization, A.O. and N.S.; methodology, A.O.; software, A.O.; validation, N.S. and A.O.; writing—review and editing, N.S. and A.K.; project administration, A.K. and M.R. and A.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available in a publicly accessible repository. The data presented in this study are openly available at https://github.com/the21composer/XAI-Metric-Evaluator.

Acknowledgments: Authors thank International Alexander Popov's Innovation Institute for Artificial Intelligence, Cybersecurity and Communications of Saint Petersburg Electrotechnical University "LETI" for support in work and research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Poursabzi-Sangdeh, F.; Goldstein, D.G.; Hofman, J.M.; Wortman Vaughan, J.; Wallach, H. Manipulating and Measuring Model Interpretability. In Proceedings of the CHI '21: CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021.
2. Tulio Ribeiro, M.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016.
3. Utkin, L.V.; Konstantinov, A.V. Ensembles of Random SHAPs. *arXiv* **2021**, arXiv:2103.03302. [CrossRef]
4. Utkin, L.V.; Konstantinov, A.V.; Vishniakov, K.A. An Imprecise SHAP as a Tool for Explaining the Class Probability Distributions under Limited Training Data. *arXiv* **2021**, arXiv:2106.09111.
5. Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
6. Doshi-Velez, F.; Kim, B. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv* **2017**, arXiv:1702.08608.
7. Rosenfeld, A. Better Metrics for Evaluating Explainable Artificial Intelligence. In Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems, Virtual Event, UK, 3–7 May 2021.

8. Hsiao, J.H.-W.; Ngai, H.H.T.; Qiu, L.; Yang, Y.; Cao, C.C. Roadmap of Designing Cognitive Metrics for Explainable Artificial Intelligence (XAI). *arXiv* **2021**, arXiv:2108.01737.
9. Keane, M.T.; Kenny, E.M.; Delaney, E.; Smyth, B. If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques. *arXiv* **2021**, arXiv:2103.01035.
10. Lin, Y.-S.; Lee, W.-C.; Berkay Celik, Z. What Do You See? Evaluation of Explainable Artificial Intelligence (XAI) Interpretability through Neural Backdoors. *arXiv* **2020**, arXiv:2009.10639.
11. Zhou, J.; Gandomi, A.H.; Chen, F.; Holzinger, A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics* **2021**, *10*, 593. [CrossRef]
12. Kim, M.-Y.; Atakishiyev, S.; Babiker, H.K.B.; Farruque, N.; Goebel, R.; Zaïane, O.R.; Motallebi, M.-H.; Rabelo, J.; Syed, T.; Yao, H.; et al. A Multi-Component Framework for the Analysis and Design of Explainable Artificial Intelligence. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 900–921. [CrossRef]
13. Vilone, G.; Longo, L. Classification of Explainable Artificial Intelligence Methods through Their Output Formats. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 615–661. [CrossRef]
14. Sarp, S.; Kuzlu, M.; Wilson, E.; Cali, U.; Guler, O. The Enlightening Role of Explainable Artificial Intelligence in Chronic Wound Classification. *Electronics* **2021**, *10*, 1406. [CrossRef]
15. Petrauskas, V.; Jasinevicius, R.; Damuleviciene, G.; Liutkevicius, A.; Janaviciute, A.; Lesauskaite, V.; Knasiene, J.; Meskauskas, Z.; Dovydaitis, J.; Kazanavicius, V.; et al. Explainable Artificial Intelligence-Based Decision Support System for Assessing the Nutrition-Related Geriatric Syndromes. *Appl. Sci.* **2021**, *11*, 11763. [CrossRef]
16. Emam, K.; Mosquera, L.; Hoptroff, R. Chapter 1: Introducing Synthetic Data Generation. In *Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2020; pp. 1–22.
17. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
18. Han, H.; Wang, W.Y.; Mao, B.H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In Proceedings of the International Conference on Intelligent Computing, Hefei, China, 23–26 August 2005; pp. 878–887.
19. Bunkhumpornpat, C.; Sinapiromsaran, K.; Lursinsap, C. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Bangkok, Thailand, 27–30 April 2009; pp. 475–482.
20. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–8 June 2008; pp. 1322–1328.
21. Douzas, G.; Bacao, F.; Last, F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Inf. Sci.* **2018**, *465*, 1–20. [CrossRef]
22. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
23. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems 27 (NIPS 2014), Montreal, QC, Canada, 8–13 December 2014; Volume 27.
24. Siddani, B.; Balachandar, S.; Moore, W.C.; Yang, Y.; Fang, R. Machine learning for physics-informed generation of dispersed multiphase flow using generative adversarial networks. *Theor. Comput. Fluid Dyn.* **2021**, *35*, 807–830. [CrossRef]
25. Coutinho-Almeida, J.; Rodrigues, P.P.; Cruz-Correia, R.J. GANs for Tabular Healthcare Data Generation: A Review on Utility and Privacy. In *Discovery Science*; Soares, C., Torgo, L., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 282–291.
26. Gupta, A.; Vedaldi, A.; Zisserman, A. Synthetic data for text localisation in natural images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2315–2324.
27. Bolón-Canedo, V.; Sánchez-Maroño, N.; Alonso-Betanzos, A. A review of feature selection methods on synthetic data. *Knowl. Inf. Syst.* **2013**, *34*, 483–519. [CrossRef]
28. Frid-Adar, M.; Klang, E.; Amitai, M.; Goldberger, J.; Greenspan, H. Synthetic data augmentation using GAN for improved liver lesion classification. In Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 4–7 April 2018; pp. 289–293.
29. Koch, B. Status and future of laser scanning, synthetic aperture radar and hyperspectral remote sensing data for forest biomass assessment. *ISPRS J. Photogramm. Remote Sens.* **2010**, *65*, 581–590. [CrossRef]
30. Wu, X.; Liang, L.; Shi, Y.; Fomel, S. FaultSeg3D: Using synthetic data sets to train an end-to-end convolutional neural network for 3D seismic fault segmentation. *Geophysics* **2019**, *84*, IM35–IM45. [CrossRef]
31. Nikolenko, S.I. Synthetic Data Outside Computer Vision. In *Synthetic Data for Deep Learning*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 217–226.
32. Pan, Z.; Yu, W.; Yi, X.; Khan, A.; Yuan, F.; Zheng, Y. Recent progress on generative adversarial networks (GANs): A survey. *IEEE Access* **2019**, *7*, 36322–36333. [CrossRef]
33. Di Mattia, F.; Galeone, P.; De Simoni, M.; Ghelfi, E. A survey on gans for anomaly detection. *arXiv* **2019**, arXiv:1906.11632.
34. Saxena, D.; Cao, J. Generative adversarial networks (GANs) challenges, solutions, and future directions. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 63.

35. Wang, Q.; Gao, J.; Lin, W.; Yuan, Y. Learning from synthetic data for crowd counting in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 8198–8207.
36. Atapour-Abarghouei, A.; Breckon, T.P. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2800–2810.
37. Liu, J.; Qu, F.; Hong, X.; Zhang, H. A small-sample wind turbine fault detection method with synthetic fault data using generative adversarial nets. *IEEE Trans. Ind. Inform.* **2018**, *15*, 3877–3888. [CrossRef]
38. Zhang, L.; Gonzalez-Garcia, A.; Van De Weijer, J.; Danelljan, M.; Khan, F.S. Synthetic data generation for end-to-end thermal infrared tracking. *IEEE Trans. Image Process.* **2018**, *28*, 1837–1850. [CrossRef] [PubMed]
39. Wang, Q.; Gao, J.; Lin, W.; Yuan, Y. Pixel-wise crowd understanding via synthetic data. *Int. J. Comput. Vis.* **2021**, *129*, 225–245. [CrossRef]
40. Chen, Y.; Li, W.; Chen, X.; Gool, L.V. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1841–1850.
41. Dunn, K.W.; Fu, C.; Ho, D.J.; Lee, S.; Han, S.; Salama, P.; Delp, E.J. DeepSynth: Three-dimensional nuclear segmentation of biological images using neural networks trained with synthetic data. *Sci. Rep.* **2019**, *9*, 18295. [CrossRef]
42. Kim, K.; Myung, H. Autoencoder-combined generative adversarial networks for synthetic image data generation and detection of jellyfish swarm. *IEEE Access* **2018**, *6*, 54207–54214. [CrossRef]
43. Torkzadehmahani, R.; Kairouz, P.; Paten, B. Dp-cgan: Differentially private synthetic data and label generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
44. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784.
45. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2015**, arXiv:1511.06434.