A STUDY COMPARING EXPLAINABILITY METHODS: A MEDICAL USER PERSPECTIVE

Miroslava MATEJOVÁ, Lucia GOJDIČOVÁ, Ján PARALIČ
Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics,
Technical University of Košice, Letná 9, 042 00 Košice, Slovak Republic, Tel. +421 55 602 4309,
E-mail: miroslava.matejova@tuke.sk

ABSTRACT

In recent years, we have witnessed the rapid development of artificial intelligence systems and their presence in various fields. These systems are very efficient and powerful, but often unclear and insufficiently transparent. Explainable artificial intelligence (XAI) methods try to solve this problem. XAI is still a developing area of research, but it already has considerable potential for improving the transparency and trustworthiness of AI models. Thanks to XAI, we can build more responsible and ethical AI systems that better serve people's needs. The aim of this study is to focus on the role of the user. Part of the work is a comparison of several explainability methods such as LIME, SHAP, ANCHORS and PDP on a selected data set from the field of medicine. The comparison of individual explainability methods from various aspects was carried out using a user study.

Keywords: *explainability, explainable artificial intelligence, medical data, user perspective, user study*

1. INTRODUCTION

Machine learning and artificial intelligence (AI) systems have become an integral part of solving complex problems in healthcare, banking and other industries. However, with the growing influence of AI on decision-making processes and critical systems, it becomes increasingly important to understand the workings and logic of these models. The opacity and lack of explainability of artificial intelligence models prevents them from trusting and understanding their decisions. For this reason, explainable artificial intelligence is becoming a key area of AI research.

Questions related to the trustworthiness and ability to interpret systems with the use of artificial intelligence are of interest to the users themselves. How can we ensure that users understand the decisions and recommendations of these systems? What are the consequences of misunderstanding and distrust in artificial intelligence? It is important to realize that for the adoption and success of artificial intelligence, it is essential that users have confidence in these systems. XAI aims to make AI models transparent and understandable to humans. This means that people should be able to understand how AI models work, what data they use, and the reasons behind their decisions.

It is difficult to mathematically define explainability. Explainability [1] can be considered an active characteristic of a model, denoting any action or procedure that a model performs with the intention of clarifying or detailing its internal functions.

2. STAKEHOLDERS AND DESIDERATA

Understanding black box machine learning models is desirable for several reasons. Though everyone has a different motivation, they are all drawn to looking within. Developing trust, identifying prejudice, and debugging models are a few examples of high-level utilization [2]. This shows that different groups of people have varying interests in the explainability of AI systems: those who use the systems and try to make them better; those who are affected by decisions made using the AI's outputs; and those who either deploy the systems for routine tasks or create the laws governing their use.

These people with different roles are commonly called stakeholders. A number of studies have been created to examine the needs of stakeholders. Similar to Arrieta et al. [1] and Langer et. al. [3] we distinguish five basic classes of stakeholders: users, developers, affected parties, deployers, and regulators. Fig. 1 shows the simple relationships between these classes.

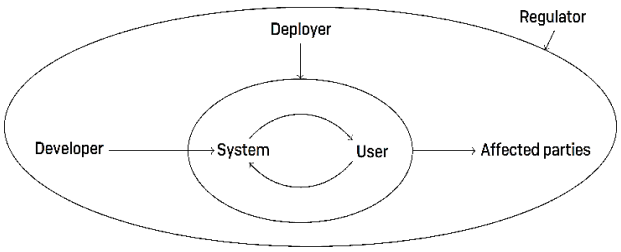

Fig. 1 The various stakeholder types connected to artificial systems and their interactions [3]

There is diversity in the desires that stem from the five groups of stakeholders. An exemplary list of desiderata, their descriptions and stakeholders holding these desiderata is presented in Table 1. The majority of users lack technical expertise and familiarity with the internal operations of the systems they utilize. The two most important things they require are usability and trust. The systems are well-versed in by developers, who are also motivated to improve them. Affected parties are those who fall inside the system's sphere of impact, and justice and morality/ethics are what matter most to them. The systems that deployers select for implementation are those that they hope will be widely adopted. Moral and legal standards are being set by regulators for the development, implementation, and general use of AI systems. In this study, we will focus on the class of users and affected parties and their desiderata.

Table 1 Exemplary stakeholders' desiderata [3]

Desideratum	Description	Stakeholder
Acceptance	Improve acceptance of systems	Deployer, Regulator
Accountability	Provide appropriate means to determine who is accountable	Regulator
Confidence	Make humans confident when using a system	User
Debugability	Identify and fix errors and bugs	Developer
Fairness	Assess and increase a system's (actual) fairness	Affected parties, Regulator
Legal Compliance	Assess and increase the legal compliance of a system	Deployer
Morality/Ethics	Assess and increase a system's compliance with moral and ethical standards	Affected parties, Regulator
Transparency	Have transparent systems	Regulator
Trust	Calibrate appropriate trust in the system	User, Deployer

3. XAI METHODS

This chapter is devoted to specific XAI methods that allow us to understand how machine learning models work. The methods that were used in the user study, namely LIME, SHAP, ANCHORS and PDP, are described in more detail.

3.1. Local interpretable model-agnostic explanations (LIME)

LIME [4] is a method that explains specific predictions of a model by building a simpler model around that prediction. It is model agnostic, meaning it can be applied to any machine learning model. It locally approximates the predictions of the original model around specific instances by creating simpler, interpretable models. The process involves generating perturbed data around a selected instance, applying the original model to that data to obtain predictions, and then weighting those predictions according to their similarity to the original instance.

3.2. Shapley additive explanations (SHAP)

The SHAP method can be used both locally and globally. It evaluates the contribution of each feature to model prediction by considering all possible combinations of features [5]. It offers several types of visualizations, such as summary plots, which show the importance and influence of features on the model, or dependency plots, which show relationships between features. There are different variants of SHAP, the most popular being Kernel SHAP, which is general but can be slower, Tree SHAP, which is specifically designed for models such as decision trees, random forests, gradient boosting, and is more efficient, and Deep SHAP, which is faster and more accurate than Kernel Shap, but only works with deep learning models.

3.3. High-precision model-agnostic explanations (ANCHORS)

This method identifies "anchors" (i.e., a set of rules or properties) that sufficiently explain the behaviour of the model for specific predictions [6]. This approach is useful in providing precise and easily interpretable explanations for machine learning decisions. It allows users to understand the conditions under which the prediction is reliable. Like LIME, Anchors are focused on local interpretability, meaning that it focuses on explaining a particular prediction or instance, rather than the entire model. Anchors can be applied to various types of models, including classifiers based on neural networks, decision trees, and more.

3.4. Partial dependence plots (PDP)

Partial Dependence Plots provide a visual way to understand the influence of one or two independent variables on a model's prediction [7]. It shows the relationship between the selected independent variable (or variables) and the dependent variable (target variable). They graphically show how the predicted mean value of the target variable changes with different values of the independent variable. It is applicable to various kinds of machine learning models, including regression models, tree models, and even some types of deep learning. Unlike methods such as LIME or SHAP, which provide a local interpretation, PDPs provide a global interpretation of the model.

4. EVALUATION METHODS

At this time, there isn't a consensus on what makes an excellent explanation or how to evaluate its validity and reliability. Techniques for assessing these attributes and instruments for examining explanations from many angles have been created within different methods of research. The study's authors [8] distinguish between two categories of criteria for measuring explained artificial intelligence:

- **Subjective metrics**: measure the degree to which individuals comprehend and interpret explanations. Are known as human-centred metrics.
- **Objective metrics**: numerical evaluations that gauge the accuracy and consistency of explanations.

We chose to employ five specific dimensions of generally acknowledged goodness of explanation: understandability, usefulness, trustworthiness, informativeness, and satisfaction, in line with the assessment methodology suggested by Hoffman et al. [9].

These five can be used for subjective evaluation by posing the following queries, all of which have been included in numerous research projects centred around XAI evaluation [9] – [12]:

- **Understandability**: From the explanation, does the user understand how the model makes a decision?
- **Usefulness**: Is the explanation useful to the user, to make better decisions or to perform an action?
- **Trustworthiness**: Does the explanation increase the user's trust in the model?
- **Informativeness**: Does the explanation provide sufficient information to explain how the model makes decisions?
- **Satisfaction**: Does the explanation of the model satisfy the user?

A user's rating on a 5-point Likert scale indicates whether they agree or disagree with the perceived fulfilment of each criterion. In our study, we used all of these five dimensions with given questions to the users.

5. RELATED WORK

In the paper [13], the authors conducted experiments where they analysed the human satisfaction and usefulness of XAI techniques such as SHAP, LIME and Permutation Importance from the user's perspective. The target group of users included domain experts, artificial intelligence experts and ordinary users, i.e. laymen. They applied the mentioned techniques to two different domains: medicine (cervical cancer detection) and sports (match player selection). In order to achieve the goal, a qualitative approach was chosen in the mentioned work. Respondents answered the question: To what extent did the XAI model help you understand the result? For domain experts in medicine (doctors) and AI experts, the SHAP and LIME methods dominated, for laymen it was the LIME method. Another task was to select a model that helped users better understand the result of the model. For doctors, the most common choice was the permutation importance method, for AI experts and laymen it was the LIME method.

In their study [14], Wang and Yin focused on finding out how do different types of explanation (feature importance, feature contribution, nearest neighbours and counterfactuals) impact people's understandings of an AI model and how do different types of explanation change people's ability of calibrating their trust in an AI model. The results of the research on 782 participants using tasks and questionnaires were that both the feature importance and counterfactual explanations increase participants' objective understanding of the model and all four types of explanations increase participants' subjective understanding of the model predicted recidivism. Their results suggest that both the feature importance and feature contribution explanation appear to help participants slightly increase their appropriate trust and decrease their undertrust in the model predicted recidivism.

In the study [15], the author investigated the effect of explainability on user trust and attitudes towards AI in terms of user interpretability and comprehensibility. 350 participants from online platforms and local universities were included in the study. Various metrics such as transparency, accountability, fairness, explainability, performance, trust and causability were also examined from the point of view of the relationship between them. The study found that factors such as fairness, accountability and transparency together affect user trust.

The mentioned studies point to the importance of investigating the impact of XAI methods on better understanding and increasing trust by different types of stakeholders. There is a clear connection between individual requirements, but it is necessary to examine several specific methods in a specific use case and type of stakeholders. A comparison of local and global methods will also be beneficial.

6. USER STUDY

In this user study, we focused on testing and comparing the XAI methods in order to get user feedback. We conducted the user study on two target groups, namely students and doctors. Computer science students and doctors combine technical and professional knowledge with practical experience. Students can assess the technical aspects, while clinicians can verify the clinical relevance and clarity of the explanations. Together they can provide a comprehensive view. In the study, 25 students and 5 doctors took part.

The testing was carried out through an online questionnaire to which selected users were invited. Before starting the testing, the participants were given instructions for filling out the questionnaire.

6.1. Dataset

The dataset used for the user study is the Pima Indians Diabetes Database, which is used in the field of gestational diabetes research. The data set contains diagnostic parameters and measurements, with the help of which the patient can be identified in time. It contains 768 entries, all participants are female and at least 21 years old. Of the total number of records, 268 records are identified as diabetic and 500 records are identified as non-diabetic. The dataset consists of 9 attributes, one of which is the target attribute. None of the attributes were deleted, nor were attributes created or merged.

6.2. AI Models

We divided the training and testing set in a ratio of 70 to 30. The data has been suitably pre-processed. Attribute values were standardized to have a mean of zero and a standard deviation of one. We replaced missing values for the attributes Glucose, BloodPressure, SkinThickness, Insulin and BMI with the mean or median. Decision tree, logistic regression, random forest and neural network (Multi-layer Perceptron classifier) models were created. The best results were achieved by the neural network, with a precision of 78%. Explainability methods were deployed for classification using this model.

Table 2 Description of PIMA dataset attributes

Attr. no.	Attribute	Attribute description	Value
1	Pregnancies	Number of pregnancies	0-17
2	Glucose	Plasma glucose concentration	44-199 (mg/dl)
3	Blood Pressure	Diastolic blood pressure	24-122 (mmHg)
4	Skin Thickness	Triceps skin fold thickness	7-99 (mm)
5	Insulin	Two-hours serum insulin	14-846 (mu U/dl)
6	BMI	Body mass index	18.2-67.1 (kg/m3)
7	Diabetes Pedigree Function	Diabetes pedigree function	0.08-2.42
8	Age	Age of an individual	21-81
9	Diabetes	Diabetes/No Diabetes	1/0

6.3. Questionnaire design

As already mentioned, part of the user study was an online questionnaire[1] that collected feedback from users. The questionnaire was divided into five sections. The first section consisted of basic questions, where the user joined the target group and answered the question of whether they had heard of explainable artificial intelligence methods. Each other section dealt with a specific explainability method and contained the same questions (listed along the five dimensions of XAI explanations in Chapter 4). At the beginning of each section with the method, the user received accompanying text in the form of model examples and story for a better understanding of the problem and outlining the goal of using the given explainability method. The accompanying text was followed by a specific explanation (Fig. 2 and Fig. 3) and finally a description of the explanation for each method separately. Users answered these questions, and the answers to each question are used to calculate a score for each of the metrics.

In Fig. 2 shows a sample explanation using the LIME method. This explanation, which is part of the questionnaire, shows a very low probability (4%) of diabetes. The attribute values of a specific patient and their influence on the prediction are shown.

An example of the global PDP method is shown in Fig. 3. Each of the graphs shows the marginal effect features have on the predicted outcome, while the vertical lines on the x-axis show the occurrence of such values in the dataset.

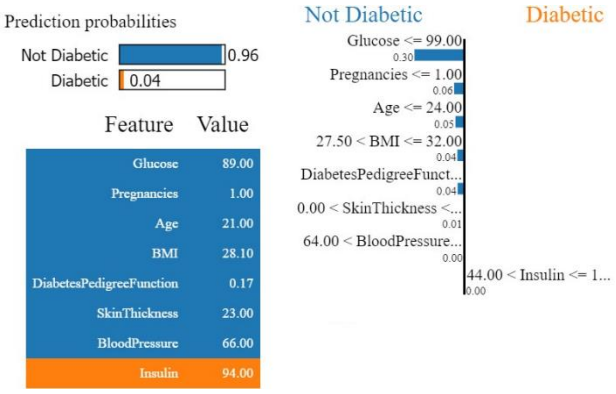

Fig. 2 Explanation using LIME in questionnaire

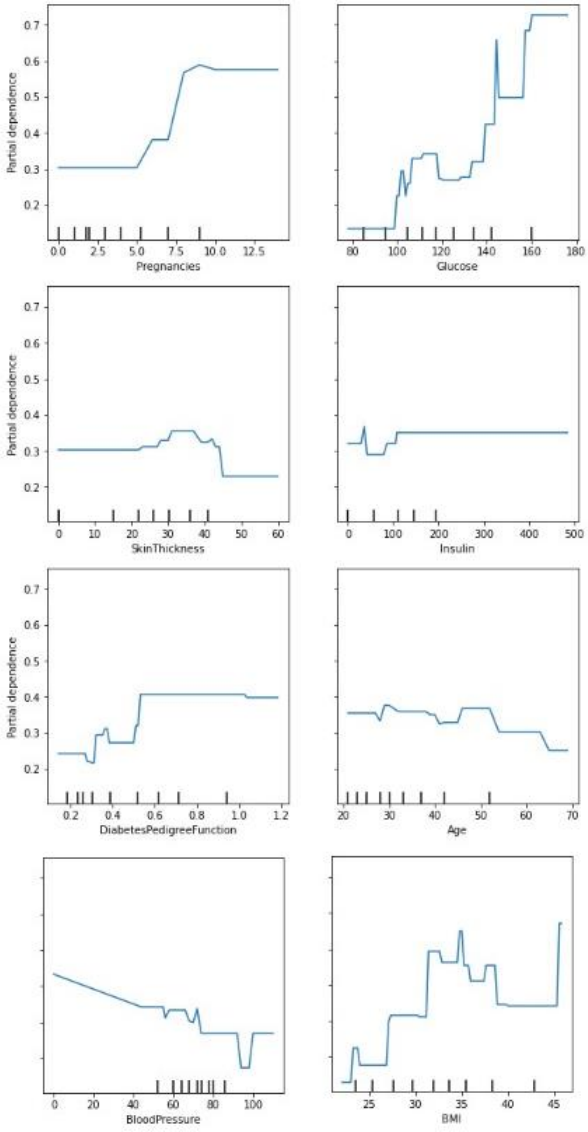

Fig. 3 Explanation using PDP in the questionnaire

[1] https://docs.google.com/forms/d/e/1FAIpQLSeXLeHwsBPHxWUf3VKnRsRBjUiYgj2p18y4bOMUNWdnIvvUxg/viewform (online questionnaire in Slovak)

6.4. User study evaluation

We used the mean score and standard deviation to evaluate the responses from the respondents in each section. We calculated the average score as the arithmetic mean of all ratings for a given question. We performed this process for each question and obtained the result for the relevant metric. The standard deviation, which is often reported along with the mean score, provides information about the variability or spread of values around the mean. A smaller standard deviation means that ratings are more consistent and centred around the mean, while a larger standard deviation indicates more diversity in respondents' opinions.

6.4.1. Target group: Students

Students are assumed to possess critical thinking skills, which is a basic prerequisite for any kind of assessment. Furthermore, it is assumed that students have knowledge in the field of data analytics, are familiar with the use of artificial intelligence, and could also have knowledge of XAI.

The evaluation in the group of students is shown in Table 3, where the best results are highlighted in bold. The LIME method is consistently rated superior to the other methods in this study according to the criteria of understandability, usefulness, informativeness and satisfaction. However, the SHAP method excels in trustworthiness, but for the complexity metric (the lower the score, the better) we notice a high average score, which in this case means that the respondents identified the method as complex and would need more knowledge. The lowest average score for understandability was achieved by the PDP method. For the usefulness metric, it was the SHAP and ANCHORS metrics. Students found the SHAP method the most difficult. The least satisfactory for them was the ANCHORS method. High standard deviations in some cases indicate greater variability in evaluations, which could point to subjective differences in interpretations of explanations between individual respondents.

Table 3 Average score and variance in a target group of students

	LIME	SHAP	ANCHORS	PDP
Understandability	**4.2±0.91**	3.76±0.97	3.76±0.72	3.68±1.18
Usefulness	**3.88±0.73**	3.6±0.96	3.6±0.87	3.76±0.88
Trustworthiness	3.96±0.89	**4.04±0.79**	3.72±0.89	3.64±0.95
Informativeness	**2.8±1.0**	3.2±1.22	2.96±0.98	3.08±1.15
Satisfaction	**4.0±0.91**	3.68±0.95	3.44±1.03	3.84±1.11

Considering the local and global methods, LIME, SHAP and ANCHOR together are rated on average as relatively understandable, trustworthy, and more informative, which indicates that respondents are comfortable with specific explanations. The global PDP method shows better satisfaction and usefulness results compared to the local methods, which could mean that the respondents appreciated its ability to provide explanations of the whole model. With this evaluation, we wanted to find out if the target group prefers local or global methods. P-values for Welch's t-test do not indicate statistical significance, indicating that there are no significant differences between the local and global methods from the students' point of view.

6.4.2. Target group: Doctors

Doctors from various specialties participated in the user study. It is predicted that doctors will have a harder time understanding the AI model but will have a much better understanding of the medical data set. We used the mean score and standard deviation for evaluation, just as in the case of the target group of students.

The average values are shown in Table 4. The best values are observed for the LIME method in three metrics, in the case of the PDP method in two metrics and once for SHAP. LIME is rated as the most understandable, while SHAP has a slightly lower score. The LIME method is also rated the highest in the satisfaction metric. The evaluation of informativeness is relatively balanced, the LIME and SHAP methods turned out better. According to doctors, the most reliable and useful method is PDP. However, the PDP method performed the worst in the understandability and informativeness metrics. In other metrics, the ANCHORS method performed worst. The standard deviations are relatively high in all metrics for all methods, indicating a wide variability in the perception of each method among different respondents.

Considering local and global methods for the target group of doctors, local methods have a better score in the metrics of understandability and informativeness. Local and global methods performed equally in the satisfaction metric. The global method fared better in both trustworthiness and usefulness. Overall, the group of doctors evaluated individual methods in all metrics with a lower score compared to the group of students.

Table 4 Average score and variance in a target group of doctors

	LIME	SHAP	ANCHORS	PDP
Understandability	**4.0±1.22**	3.8±1.1	3.4±1.52	3.2±1.64
Usefulness	2.8±1.3	2.8±0.84	2.6±1.34	**3.0±1.22**
Trustworthiness	3.8±0.84	3.8±0.84	3.0±1.0	**4.0±0.71**
Informativeness	**3.2±1.3**	**3.2±1.3**	3.6±1.52	3.6±1.52
Satisfaction	**3.0±1.41**	2.8±1.3	2.6±1.52	2.8±0.84

7. CONCLUSIONS

In this study, we focused on analysing the explainability of artificial intelligence models from the users' perspective.

We analysed various existing solutions and experiments with explainability methods and artificial intelligence models. We conducted a user study in which two target groups were included. 30 respondents took part in the user study, including 25 students and 5 doctors.

For the student target group, the LIME method was rated the best in terms of understandability, usefulness, informativeness and satisfaction, while SHAP excelled in trustworthiness. Further evaluation showed that none of the XAI methods is a one-size-fits-all solution and respondents' preferences varied. High standard deviations in some metrics indicate variability in the perception and interpretation of explanations between individual students. There are no statistically significant differences between local and global methods.

In the focus group doctors, we found that they prefer methods that provide clear and direct explanations, with the LIME method standing out in terms of understandability, informativeness and satisfaction, while the PDP was rated as the most reliable and useful. The results indicated that there is considerable variability in the perception of different XAI methods among doctors. Despite assumptions that local and global XAI methods might be perceived differently, the study found no significant differences. The recommendations and suggestions of doctors resulted in the need to simplify explanations and interaction with XAI applications to make them more intuitive and better adapted to the needs of users in the medical field. They emphasize the importance of simple, seamless interfaces that enable quick and reliable retrieval of the required information without the need for technical knowledge.

All these findings from the user study point to the importance of tailoring XAI tools to the needs and preferences of end users in order to maximize their potential and promote their acceptance and effective use in practice.

ACKNOWLEDGMENTS

This work was supported by the Slovak Research and Development Agency under the contracts No. APVV-22-0414 and APVV-20-0232, and by the Scientific Grant Agency of the Ministry of Education, Research, Development and Youth of the Slovak Republic under grants No. 1/0259/24 and 1/0685/21.

REFERENCES

[1] ARRIETA, A.B. – DÍAZ-RODRÍGUEZ, N. – DEL SER, J. – BENNETOT, A. – TABIK, S. – BARBADO, A. – GARCIA, S. – GIL-LOPEZ, S. – MOLINA, D. – BENJAMINS, R. – CHATILA, R. – HERRERA, F.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, vol. 58, pp. 82–115, 2020.

[2] BRENNEN, A.: What do people really want when they say they want ‘explainable AI?’ we asked 60 stakeholders, *Conference on Human Factors in Computing Systems - Proceedings*, Apr. 2020.

[3] LANGER, M. – OSTER, D. – SPEITH, T. – HERMANNS, H. – KÄSTNER, L. – SCHMIDT, E. – SESING, A. – BAUM, K.: What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artificial Intelligence, vol. 296, 2021.

[4] RIBEIRO, M. T. – SINGH, S. – GUESTRIN, C.: “Why Should I Trust You?”: Explaining the Predictions of Any Classifier, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

[5] LUNDBERG, S. M. – LEE, S. I.: A Unified Approach to Interpreting Model Predictions, *Advances in Neural Information Processing Systems*, pp. 4766–4775, Dec. 2017.

[6] RIBEIRO, M. T. – SINGH, S. – GUESTRIN, C.: Anchors: High-Precision Model-Agnostic Explanations, *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[7] FRIEDMAN, J. H.: Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[8] ROSENFELD, A.: Better Metrics for Evaluating Explainable Artificial Intelligence, *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 45–50, 2021.

[9] HOFFMAN, R. R. – MUELLER, S. T. – KLEIN, G. – LITMAN, J.: Metrics for Explainable AI: Challenges and Prospects. *CoRR*, 2018.

[10] CHROMIK, M. – SCHUESSLER, M.: A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI. *ExSS-ATEC@IUI*, 2020.

[11] DIEBER, J. – KIRRANE, S.: Why model why? Assessing the strengths and limitations of LIME. *ArXiv*, 2020.

[12] AECHTNER, J. – CABRERA, L. – KATWAL, D. – ONGHENA, P. – VALENZUELA, D. P. – WILBIK, A.: Comparing User Perception of Explanations Developed with XAI Methods, *IEEE International Conference on Fuzzy Systems*, July 2022.

[13] DAUDT, F. – CINALLI, D. – GARCIA, A. C. B.: Research on Explainable Artificial Intelligence Techniques: An User Perspective, *IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pp. 144–149, 2021.

[14] WANG, X. – YIN, M.: Effects of Explanations in AI-Assisted Decision Making: Principles and Comparisons, *ACM Transactions on Interactive Intelligent Systems*, vol. 12, no. 4, 2022.

[15] SHIN, D.: The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, vol. 146, 2021.

Received June 28, 2024, accepted November 8, 2024

BIOGRAPHIES

Miroslava Matejová graduated (MSc) in 2022 at the Department of Cybernetics and Artificial Intelligence of the Faculty of Electrical Engineering and Informatics at Technical University in Košice. After that, she continues at the department as a PhD student and her scientific research focuses on data analytics, machine learning and explainable artificial intelligence in different domains.

Lucia Gojdičová graduated (MSc) in 2024 at the Department of Cybernetics and Artificial Intelligence of the Faculty of Electrical Engineering and Informatics at Technical University in Košice.

Ján Paralič graduated (MSc) in 1992 at the Department of Cybernetics and Artificial Intelligence of the Faculty of Electrical Engineering and Informatics at Technical University in Košice. He defended his PhD in the field of Artificial Intelligence in 1998. In 2004 he was promoted to Associate Professor in Artificial Intelligence from the Technical University of Košice, and in 2011 he was appointed Full Professor in Business Information Systems. He is professor at the Dept. of Cybernetics and Artificial Intelligence, TUKE and leads the Data Science research team of researchers and doctoral students in the field of (medical) data analytics, including explainability of data analytic models.