A New XAI Evaluation Metric for Classification

Asmaa M. Elgezawy
Information Systems Department
Faculty of computers and information
Menofia University
Menofia, Egypt
Asma.elgezawy@ci.menofia.edu.eg

Hatem M. Abdul-kader
Information Systems Department
Faculty of computers and information,
Menofia University
Menofia, Egypt
Hatem.abdelkader@ci.menofia.edu.eg

Asmaa H. Elsaid
Information Systems Department
Faculty of computers and information,
Menofia University
Menofia, Egypt
Asmaa.elsayed@ci.menofia.edu.eg

Abstract— Explainable AI (XAI) has become a hot topic across multiple sectors. In practical applications, classification models are severely constrained by the absence of transparency, which undermines trust and has a black-box nature, leading to a range of problems. Classification models necessitate the use of XAI approaches to address these limitations effectively. The Mean Evaluation of Metrics Change (MEMC) is a novel metric introduced in this research for evaluating the performance of Explainable AI techniques on a global scale, like post-hoc and intrinsic XAI for classification techniques on tabular data. The proposed MEMC metric is formed from a combination of the existing standard evaluation measures used for evaluating classification. The proposed MEMC has proven to be the convenient metric for determining the best explainer for a produced classification. The proposed MEMC metric is validated using a heart dataset from the healthcare sector. The experimental results show that the Artificial Neural Network (ANN) approach performed effectively on the heart dataset as an intrinsic XAI in machine learning. Deep Neural Network (DNN) also performs better as an intrinsic XAI technique when applied to this dataset. Furthermore, ANCHORS has shown strong performance as a post-hoc XAI technique when Random Forest (RF) and XG-Boost are used as classification models.

Keywords —Explainable AI (XAI), MEMC, Intrinsic XAI, Post-hoc XAI, ANCHORS.

I. INTRODUCTION

Artificial Intelligence (AI) is now essential to many industries, including healthcare and finance [1,2]. The improvement of AI systems has been driven by the need for flexibility, resourcefulness, inventiveness, real-time reactivity, and long-term reflection to demonstrate competence in complicated surroundings and social contexts. However, to ensure that AI systems are transparent, understandable, and trustworthy for humans, Explainable AI (XAI) has become an important area of AI research [3,4,5]. One of the significant applications of XAI is tabular data which refers to data that is structured in a tabular format, arranged in rows and columns. Tabular data is used in various domains such as credit scoring, fraud detection, healthcare, marketing, and predictive policing. Tabular XAI techniques are used to explain the decision-making process of machine learning models trained on tabular data.

Furthermore, XAI techniques can be categorized according to their scope, which pertains to whether the algorithm offers local or global explanations [4]. local explanations provide insights into why specific decisions were made by focusing on individual instances [4]. On the other hand, Global explanations provide insight into how a machine-learning model functions by demonstrating feature importance, partial dependence plots, SHAP values, and LIME weights [4,6] Global XAI techniques can be classified based on usage. Usage refers to whether the algorithm is intrinsic or post-hoc. Intrinsic means any change in the architecture will need significant changes in the method itself (Model-specific) such as decision trees or linear regression or by integrating domain-specific knowledge into the model [3]. Post-hoc, which refers to the process of elucidating a model's decisions after it has been trained such as feature attribution, rule extraction, and model-agnostic techniques such as LIME or SHAP [3,4,6]. Intrinsic explainability is considered a glass box because all details in the model are transparent and clear, while post-hoc explainability is regarded as a black box model.

Classification involves determining the category to which an observation belongs from a given set of categories. Many classification models generate black-box predictions which can be problematic, particularly in critical domains like healthcare or finance where reliable models are essential. The absence of transparency in classification techniques undermines trust and leads to a range of problems. Models like complex neural networks and machine learning algorithms make it challenging to understand the reasoning behind their decisions, which in turn hampers trust and acceptance. Additionally, even though deep neural networks (DNNs) have had considerable success recently, their usage in high-risk systems like healthcare has raised concerns due to their nature as a black box [7]. To address these issues, Explainable Artificial Intelligence (XAI) techniques have emerged with the aim of providing transparency, interpretability, fairness, and trustworthiness by generating explanations for the model's predictions. Also, classification models require XAI methods to be effectively and efficiently applied in real-world scenarios while maintaining high accuracy. These techniques, such as rule-based models, feature importance analysis, and attention mechanisms, help uncover the internal mechanisms of black box models, enhance understanding, identify biases, and promote trust and accountability in decision-making processes. Moreover, when the researcher knows the reasons for classification results, classification models have been improved.

The XAI evaluation metric can assist in selecting appropriate XAI methods for the classification model, thereby enhancing the credibility of the black-box model. Additionally, the XAI evaluation metric can aid in the selection of the most appropriate glass box models for many domains. So, classification models will be modified to improve their performance and accuracy in achieving the

optimal prediction by using the explanation of the appropriate XAI method. In the previous research papers, the evaluation metrics for XAI methods can be categorized as either human-centered or computer-centered. In this paper, Mean Evaluation of Metrics Change (MEMC) has been applied as an XAI evaluation metric on a global scope such as intrinsic or post-hoc XAI techniques for the classification of the heart domain. MEMC is computer-centered, but it includes many pre-used evaluation measures in classification together, so it has noticed good results when MEMC has been used. The focus lies on the estimation of explanation quality leads to improvement of the strengths and weaknesses associated with explanation methods.

The rest of the paper is recognized as follows: Section 2 presents the related work and its constraints on XAI evaluation. Section 3 presents a new XAI evaluation metric for classification. Section 4 presents the performance evaluation and experimental results with a discussion on global XAI techniques, such as post-hoc and intrinsic XAI techniques. Section 5, the conclusion Section of the proposed study. Finally, this paper provides guidance for future research with promising and essential directions in the evaluation of XAI techniques.

II. RELATED WORK

In this section, several research articles are presented that focus on evaluating Explainable Artificial Intelligence (XAI) methods. The goal is to determine the most effective XAI method for identifying the most influential features in a dataset, which can contribute to accurate predictions. By conducting this evaluation, the prediction model is developed and enhanced, considering the significant features that have been identified.

Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litma propose four XAI measures to evaluate the effectiveness of XAI systems: explanation goodness, explanation satisfaction, scale validation, and performance. Explanation goodness measures the clarity and precision of the explanation, while explanation satisfaction measures the degree to which users feel that they understand the AI system or process being explained to them. Scale validation measures the validity of the scale used to measure explanation satisfaction, and performance measures the user's performance after being given an explanation. In addition to these quantitative measurements, XAI techniques can also be evaluated qualitatively by surveys, where individuals are asked to complete questionnaires to gauge their understanding of the explanations provided. User studies involve participants being assigned tasks that require them to utilize the explanations for decision-making. Additionally, expert review entails seeking feedback from domain experts who assess the explanations' accuracy, usefulness, and trustworthiness. We believe that the proposed measures are important for evaluating the effectiveness of XAI systems and that our work provides a foundation for the development of more effective XAI systems [8].

p. Lopes and E. Silva present a comprehensive examination of evaluation methods for XAI systems. It is crucial to thoroughly assess the explainability components of these systems to ensure their effectiveness and usefulness for end users. Explanations provided by XAI can have a positive impact on user understanding and trust in machine learning systems. The evaluation methods for XAI systems can be categorized as either human-centered or computer-centered. In the human-centered approach, the evaluation depends on user feedback about trust, usefulness, satisfaction of explanations, understandability, and performance. The computer-centered approach, on the other hand, revolves around interpretability and fidelity. A simple metric for evaluating the interpretability of models. The metric takes into account various factors such as the number of runtime, the total number of rules in the set, overlap, and the maximum width of all the elements in it. Fidelity measures the correlation between the model's performance drop when certain features are removed and the relevance scores (attributions) assigned to those features. Additionally, the sensitivity of explanations implies that when an input and baseline vary in one feature and yield different predictions, a non-zero attribution should be allocated to that differing feature [9]. This paper doesn't mention how XAI evaluation methods would be used to select an XAI method for a particular application.

Isaac Ronald Ward and Ling Wang in their study, utilized machine learning (ML)-based feature importance methods to gain insights into the contribution of specific features to the predictions made by the model. They employed traditional methods such as Mean Decrease of Impurity (MDI) and Mean Decrease in Accuracy (MDA), as well as explainable artificial intelligence (XAI) methods like LIME and SHAP. MDI calculates feature importance by considering the depth of nodes within decision trees of tree-based models. Features used by nodes closer to the top of the decision tree are deemed to have a more significant impact on the final decision compared to those used at lower levels. On the other hand, MDA measures the decrease in model accuracy on the test set when a particular feature is randomized or permuted. The magnitude of the accuracy drop indicates the model's reliance on that feature, serving as an estimate of feature importance. These methods were applied to Random Forest, Extra Trees Extreme, and Gradient Boosting classifiers to detect feature contributions. The results showed that both XAI methods, LIME and SHAP, were able to successfully identify important and unimportant features. However, SHAP slightly outperformed LIME in this regard. The evaluation of the feature scoring methods' validity was carried out by confirming their capability to assess the significance of features that held a certain level of known importance. There are limitations such as the signals in these experiments are only derived from dispensing data; there are no adverse event signals that can be linked from a joint database [10].

Y. Zhang and F. Xu proposed a new XAI evaluation metric called MDMC. It is specifically centered around regression models, and its primary focus in this paper is to compare the performance of LIME and SHAP through the establishment of the XAI evaluation metric - MDMC. Furthermore, MDMC has also been employed in the context of random perturbation. The findings indicate that LIME is a better fit for ANN models and Random Forest models based on bagging algorithms, while SHAP is better suited for Light-GBM models based on boosting algorithms. MDMC incorporates three commonly used metrics in regression, namely Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R2). These metrics are widely used to evaluate the performance of regression models [1].

Avi. Rosenfeld proposes that user studies are commonly used to evaluate XAI; however, studies that neglect objective metrics might lack significance and be prone to confirmation bias, especially when relying on unnecessary low-fidelity explanations. To tackle this issue, the paper suggests a transformative approach to evaluate XAI. This approach revolves around metrics that quantitatively measure both the explanation itself and how well it aligns with the XAI objective. Four specific metrics are suggested, including D, which captures the performance differences between the explanation's logic and the agent's actual performance, R, the number of rules produced by the explanation, F, the number of features employed in generating the explanation, and S, the stability of the explanation. The authors assert that user studies incorporating these metrics into their assessments are intrinsically more reliable and should be integrated into future XAI research [11].

Measuring the correctness of interpretability in XAI poses difficulties due to current evaluation methods relying on subjective human input or involving high computational costs for automated evaluation. However, using human subjective measurements in evaluations can be time-consuming, tedious, and prone to introducing biases and inaccuracies [7]. In addition to user bias, a questionnaire has been conducted among expert users, revealing that they hold different viewpoints about important features compared to others. This difficulty in obtaining reliable and consistent results from human evaluation studies arises as a result. Additionally, the complex nature of explainability makes it challenging to devise quantitative metrics, which adds to the difficulty of objectively measuring and evaluating XAI techniques. Also, one of the main challenges in evaluating XAI techniques lies in the absence of standardized evaluation metrics, hindering the comparison and assessment of different methods to determine their relative effectiveness.

In classification reports, recall functions as an indicator of the classifier's capability to correctly identify all instances of positive samples, while accuracy quantifies how accurately a model predicts outcomes for a given input. To comprehensively evaluate a model's efficacy, the consideration of both precision and recall is imperative; these metrics assess the performance of classification models. Furthermore, specificity is derived by dividing the total count of true negatives by the overall count of actual negatives. These metrics are widely comprehended and employed, rendering them invaluable for comparing the performance of different eXplainable Artificial Intelligence (XAI) methods. These metrics can be used to appraise the accuracy, interpretability, and fairness of XAI explanations. While recall and precision are applicable for measuring the accuracy of explanations, specificity can be employed to measure the fairness of these explanations. We are inspired by all the previously mentioned research, to conduct an evaluation of XAI metric for classification by using various performance measures. From the idea of MDA structure, besides the calculation of accuracy, we also calculate precision, recall, and specificity as a set of metric evaluations of classification that are widely recognized and commonly used to evaluate the performance of a predictive classification model. Then merge all measures and apply all of them to the test data after deleting the most important features resulting from the XAI techniques to determine which feature has a greater impact on prediction results. Then based on that, it is determined which of the XAI is better for this prediction model. we apply this new evaluation metric on intrinsic XAI such as Random Forest, XG-Boost, Logistic Regression, Decision tree, and EBM and apply it on post-hoc XAI such as SHAP, LIME, and Anchor.

III. XAI Evaluation Metric

Researchers commonly rely on several metrics to evaluate a model's performance in classification tasks. The confusion matrix is constructed by comparing the predicted outcomes with the real values. It contains information that helps determine how well the classifier performs. Diverse performance metrics, including accuracy, precision, recall, and specificity, can be employed to assess the effectiveness of the classification model and provide a more comprehensive evaluation of the model's performance in such situations. A combination of these metrics can offer a complete understanding of the model's strengths and weaknesses. This paper creates a new evaluation metric that includes many evaluation metrics together. It is applied to intrinsic XAI and Post-hoc XAI in the global scope to determine the best XAI technique that is used to obtain the most important features.

XAI Evaluation metric: Mean Evaluation of Metrics Change (MEMC)

To evaluate the performance of specific classification models, various metrics such as accuracy, precision, recall, and specificity can be calculated. Following this, XAI techniques can be applied to provide insights into the underlying decision-making processes of these models. After applying XAI techniques whether post-hoc or Intrinsic, retrieve the most important features. First, delete the most important feature from the original dataset and calculate new metrics (accuracy, precision, recall, specificity), and second, delete the most two important features from the original dataset and calculate new metrics (accuracy, precision, recall, specificity). Third, delete the three most important features from the original dataset and calculate new metrics (accuracy, precision, recall, specificity). The three most important features that are retrieved are not necessarily to be the same when applying different XAI techniques. To identify the optimal XAI technique for given classification models, the Mean Evaluation of Metrics Change (MEMC) can be computed. The XAI technique with the highest MEMC value is expected to have the most significant impact when removing the three most important features. Consequently, it is considered the most appropriate XAI technique for the given classification models.

In Perturb the original data is according to the output of the XAI method, and input the modified data into the prediction model to obtain new metrics (accuracy*, precision*, recall*, specificity*), then the degree of Evaluation change of the metrics (D) can be is defined as:

$$D = f(M - M^*) \qquad (1)$$

M is the original metric, and M* is the changed metric. Combining the degree of change of all metrics, the final evaluation metric (MEMC) can be defined as:

$$\text{MEMC} = \frac{1}{n}\sum_{i=1}^{n} D = \frac{1}{n}\sum_{i=1}^{n} f(M - M^*) \qquad (2)$$

$$= \frac{1}{n}\sum_{i=1}^{n}[(\text{accurcy} - \text{accuracy}^{*}) + (\text{precision} - \text{precision}^{*}) + (\text{recall} - \text{recall}^{*}) + (\text{specificity} - \text{specificity}^{*})]$$

It should be noted that the values of accuracy, precision, and recall in equation (2) need to be used after normalization. In theory, the larger the value of MEMC, it means that the black-box prediction model has made significant changes to the data set, thus proving the effectiveness of XAI.

IV. Performance Evaluation

A. Dataset Description

Detecting heart disease early can be a challenging task, despite it being a significant global cause of death. Machine learning and deep learning models have shown promise in accurately predicting and diagnosing heart disease [12]. The Cleveland Clinic Foundation provides a heart dataset with 14 columns and 303 rows, containing patient characteristics such as age, sex, chest pain type, and other medical attributes. The Heart dataset has a slight class imbalance, with 54% of examples showing no signs of heart disease and 46% showing signs of it. The k-fold cross-validation technique proves to be beneficial in addressing such imbalances. This method constructs folds that maintain a similar class distribution as the original dataset, reducing the risk of overfitting to the majority class and providing more accurate performance estimates for the model.

Dataset source: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

B. Black-box Model Predictions

ML techniques have made significant advances in the medical field, including the use of algorithms such as Random Forest, Decision Tree, XG-Boost, Logistic Regression, KNN, and ANN. Additionally, DNN is used as Deep Learning in this field. To evaluate the model's performance, compute (accuracy, precision, recall, and specificity). The results of each model's performance as shown in Table 1.

C. Experiment Result and Discussion

In this section, MEMC is applied for testing XAI techniques such as post-hoc XAI techniques like (Random Forest, XG-Boost, Logistic Regression, Decision tree, EBM, ANN, and DNN) and intrinsic XAI techniques like (LIME, SHAP, and ANCHORS) for classification to detect which of them is the best technique. Figure 1 depicts all steps in the experiment in order.

TABLE 1. Classification performance results

Model	Measures			
	Accuracy	Precision	Recall	Specificity
Random Forest	98.4	98.8	98.3	98.6
XG-Boost	98.2	97.8	98.8	97.6
Logistic Regression	85.7	83.0	91.5	79.2
Decision Tree	97.1	96.6	98.0	96.2
KNN	85.2	84.8	87.3	82.8
ANN	93.6	92.3	95.0	92.3
EBM	97.1	97.8	96.8	97.5
DNN	95.0	94.8	95.4	94.7

Explaining with Intrinsic XAI Techniques

By applying intrinsic XAI techniques, including Random Forest, XG-Boost, Logistic Regression, Decision tree, EBM, ANN, and DNN, the three most significant features are retrieved and are presented in Table 2. Based on the impact of each feature on prediction outcome, which has been obtained from previous algorithms as shown in Table 2, the MEMC metric has been executed. First, delete the first most important feature according to results in Table 2, then compute accuracy, recall, precision, and specificity. It is observed that all values of standard evaluation measures may be different from the original results which are shown in Table 1. As well when deleting the second and third most important features, all values of standard evaluation measures may be different from the original results that have been computed without any deletion as shown in Table 3. Finally, normalize all values of accuracy, recall, precision, and Specificity then compute the MEMC metric as shown in equation 2. The results of MEMC for each glass box have been shown in Table 4. The equation of MEMC contains many standard performance measures in the classification, which distinguishes it from other evaluation metrics. The MEMC is used to determine which of these intrinsic XAI techniques is better in finding the most important features that have the most impact in many standard evaluation measures together. It is observed that when the value of MEMC increases, this means this XAI technique is the best for obtaining the most important features that are used for prediction by classification models. Table 4 displays the experimental results indicating that DNN, ANN, Decision Tree, Random Forest, and EBM are the top performing intrinsic XAI techniques.

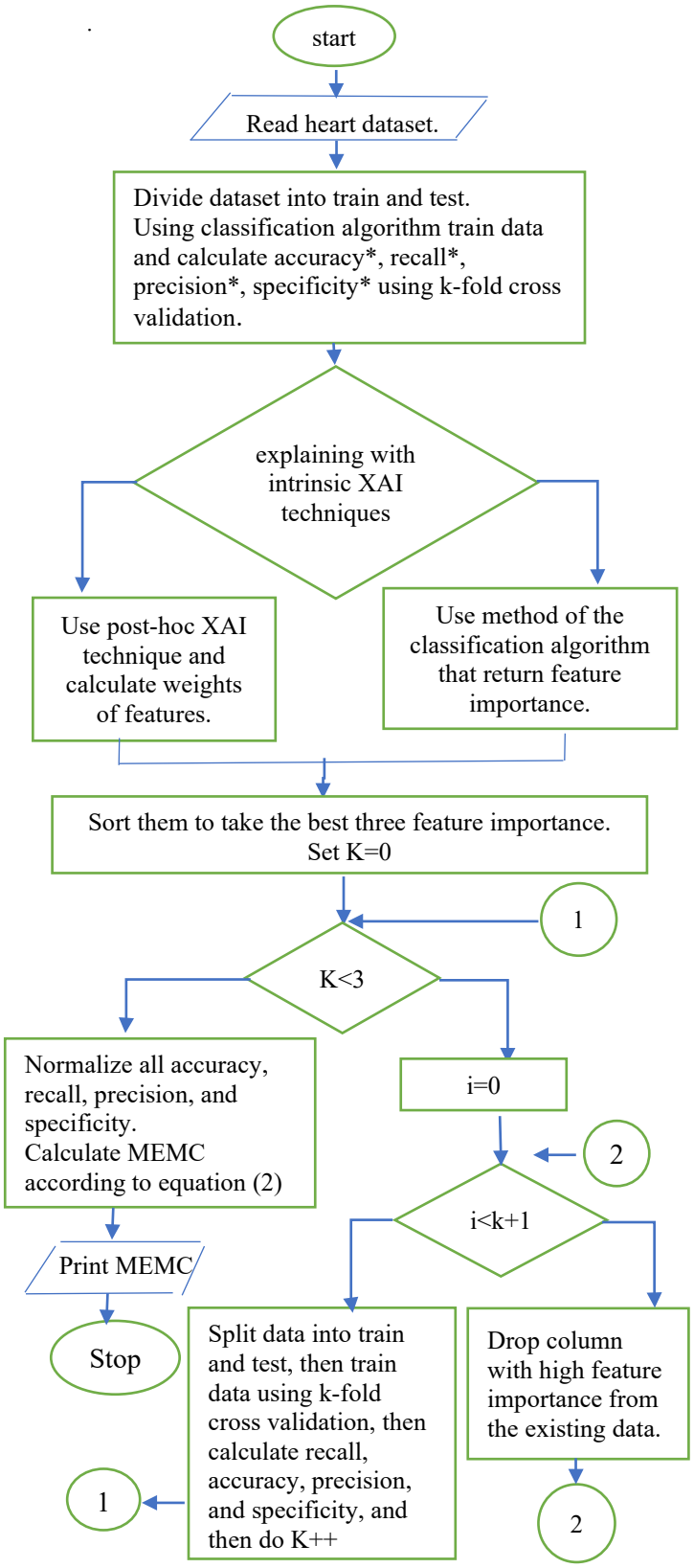

Fig. 1. The generic Flow of the experiment

TABLE 2. Features importance of intrinsic XAI techniques

Model	The three most features importance
Random Forest	['thalach','cp','thal']
XG-Boost	['cp', 'thal', 'ca']
Logistic Regression	['cp','thalach','slope']
Decision Tree	['cp','oldpeak','thal']
EBM	['ca','thal','cp']
ANN	['sex','cp','age']
DNN	['exang','sex','cp']

TABLE 3. Changes in classification performance results

Model	***Measures***			
	Accuracy	***Precision***	***Recall***	***Specificity***
Without the first most important features				
Random Forest	98.3	98.8	98.0	98.6
XG-Boost	96.5	97.8	95.5	97.5
Logistic Regression	81.5	82.2	82.6	80.3
Decision Tree	95.1	95.1	95.5	94.5
EBM	96.2	97.0	95.8	96.7
ANN	91.3	88.9	93.8	88.9
DNN	81.5	77.0	88.3	75.0
Without the two most important features				
Random Forest	97.4	98.3	96.8	98.1
XG-Boost	97.4	97.5	97.5	97.3
Logistic Regression	80.5	81.4	81.4	79.5
Decision Tree	95.6	96.0	95.5	95.6
EBM	95.4	96.1	95.3	95.6
ANN	87.9	87.6	87.7	88.2
DNN	78.8	75.4	83.8	73.9
Without the three most important features				
Random Forest	97.1	98.3	96.3	98.1
XG-Boost	96.0	96.5	95.8	96.2
Logistic Regression	79.8	80.0	82.1	77.3
Decision Tree	93.2	95.5	91.6	95.1
EBM	94.3	94.6	94.5	94.0
ANN	86.6	85.3	87.5	85.8
DNN	76.3	73.1	81.8	71.2

TABLE 4. MEMC of intrinsic XAI techniques

Model	MEMC
Random Forest	2.5396884657939958
XG-Boost	2.309770958116008
Logistic Regression	2.4722174500568497
Decision Tree	2.5974696779225437
EBM	2.527294618890463
ANN	2.8669977313368356
DNN	3.465272219094056

Explaining with Post-hoc XAI Techniques

Post-hoc explainability is the procedure of clarifying a model's decisions after it has undergone training. Post-hoc XAI techniques such as LIME, SHAP, and ANCHORS can be used in the global scope except LIME which is used in the local scope. To calculate global explanation by LIME, add all weights for each feature in all instances, then divide by the total number of instances to calculate an average of weights for each feature and sort the average weight of each feature. Finally, select the top three important features. To employ post-hoc XAI techniques, the initial step involves choosing a machine learning algorithm suitable for classification, in which the most important features are to be determined.

After examining the classification performance results presented in Table 1, it was determined that Random Forest, XG-Boost, and EBM exhibit the most favorable performance in terms of evaluation measures such as accuracy, recall, precision, and specificity for the heart domain. These algorithms were thus chosen as the best candidates for further analysis. First, apply LIME, SHAP, and ANCHORS after the RF classifier. The three most important features are obtained and shown in Table 5. To compute the MDMC equation, the first most important feature is deleted, and new values of accuracy, recall, precision, and specificity are computed using RF. It is observed that these values are different from the original values in Table 1. The process is then repeated for the two most important features and again for the three most important features. Normalize all values of accuracy, recall, precision, and specificity before computing the MEMC. The sum of the degree of change evaluation of the three metrics results in MEMC, which is shown in Table 5 when the three metrics are combined. Finally, it is observed that Anchors is better than LIME and SHAP in the case of applying RF as a prediction model because Anchors has the highest value of MEMC.

TABLE 5. Post-hoc XAI with RF

Model	MEMC	The most three feature importance
LIME	2.5738753245233634	['restecg','cp','oldpeak']
SHAP	2.477609811921759	['thal','ca','cp']
Anchor	2.87639238654984	['cp','ca','oldpeak']

To be able to apply LIME, SHAP, and ANCHORS after XG-Boost, repeat the same steps as before, but with XG-Boost algorithm instead of applying it to RF algorithm. It is observed that Anchor is better than LIME and SHAP in the case of applying XG-Boost as shown in Table 6. This means that 'cp', 'ca', and 'oldpeak' features in order have the highest impact in controlling the prediction result than other features.

TABLE 6. Post-hoc XAI with XG-Boost

Model	MEMC	The most three feature importance
LIME	0.6137865908073462	['restecg','cp','oldpeak']
SHAP	1.6272410455986845	['thal','ca','cp']
Anchor	1.9988550481603695	['cp','oldpeak','ca']

When using LIME, SHAP, and ANCHORS after prediction using EBM, the most important features are obtained. These important features are used to compute MEMC by repeating the previous steps used in the case of RF. The most important features and the results of MEMC for EBM are shown in Table 7. It is observed that SHAP takes a long time for execution and observe LIME and SHAP the best Choice XAI techniques after prediction by EBM for the heart domain.

TABLE 7. Post-hoc XAI with EBM

Model	MEMC	The most three feature importance
LIME	3.1177996990582564	['cp','exang','ca']
SHAP	3.1387695891126897	['cp','exang','ca']
Anchor	2.527294618890463	['ca','thal','cp']

V. CONCLUSION

This paper demonstrates how the proposed metric enables easy evaluation of global explanations for classification on real datasets such as the heart dataset and facilitates the characterization of the quality of global explanation methods such as post-hoc XAI and intrinsic XAI. Based on the implemented techniques, it is evident that ANN, Decision Tree, and Random Forest as ML techniques perform exceptionally well as glass box models for explaining the classification results of heart disease using the experimental data. Deep Neural Network (DNN) also performs better as an intrinsic XAI technique when applied to this data set. Furthermore, it is observed that ANCHORS outperforms LIME and SHAP as the preferred post-hoc XAI techniques

when applied after Random Forest and XG-Boost as classification models. MEMC presents itself as a valuable and model-agnostic asset for appraising XAI methods. Its scalability and objectivity further enhance its worth, rendering MEMC a valuable tool applicable across various domains for evaluating XAI methods for classification models.

VI. FUTURE WORK

Future research will improve the proposed metric by adding time complexity to it and applying the new metric to deep learning models in many domains not only tabular data, but also image, video, and text data in classification models. It will enable the efficiency of different XAI technology to be measured more effectively. It will provide valuable insights into the decision-making processes of complex models and facilitate the creation of more transparent AI systems. Also, will focus on overcoming challenges related to high-dimensional data and complex architectures, ensuring the metric's practicality for real-world datasets. Ultimately, these efforts will enhance model interpretability and contribute to the responsible deployment of AI in various applications.

REFERENCES

[1] y. Zhang,F. Xu, J. Zou and O.L.Petrosian, "XAI Evaluation: Evaluating Black-Box Model Explanations for Prediction", 2021 International Conference on National Networks and Neurotechnologies (NeuroNT), DOI:10.1109/NEURONT53022.2021.9472817.

[2] M. SAHAKYAN, Z. AUNG AND T. RAHWAN, "Explainable Artificial Intelligence for Tabular Data: A Survey". VOLUME 9, 2021, Digital Object Identifier 10.1109/ACCESS.2021.3116481.

[3] B. Arrietaa , N. Díaz-Rodríguez , J. D. Ser, A. Bennetot, S.Tabikg , A. Barbado , S. Garcia, S. Gil-Lopez , D. Molina , R. Benjamins , R. Chatilaf and F. Herrera," Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI" , 2020 ELSEVER. Journal. Information Fusion 58 (2020) 82–115.

[4] Chaddad , J. Peng, J. Xu , and A. Bouridane,"Survey of Explainable AI Techniques in Healthcare.Sensors 2023, 23, 634. https://doi.org/10.3390/s23020634.

[5] J. Duell , X. Fan , B. Burnett , G. Aarts and S. Zhou,"A Comparison of Explanations Given by Explainable Artificial Intelligence Methods on Analysing Electronic Health Records", 2021 IEEE EMBS International Conference on Information Technology Applications in Biomedicine (ITAB).

[6] M. S. Khan, M. Nayebpour , M. Li , H. El-Amine , N. Koizumi and J. L. Olds. "Explainable AI: A Neurally-Inspired Decision Stack Framework", biomimetics, Neurally-Inspired Decision Stack Framework. Biomimetics 2022, 7, 127. https://doi.org/10.3390/biomimetics7030127.

[7] Yi-Shan Lin, Wen-Chuan Lee,and Z. Berkay Celik, "What Do You See? Evaluation of Explainable Artificial Intelligence (XAI) Interpretability through Neural Backdoors", KDD '21, August 14–18, 2021, Virtual Event, Singapore.

[8] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litma, "Metrics for Explainable AI: Challenges and Prospects", 2018, https://doi.org/10.48550/arXiv.1812.04608.

[9] Lopes, P.; Silva, E.; Braga, C.; Oliveira, T. and Rosado, L. XAI Systems Evaluation: A Review of Human and Computer-Centred Methods. Appl. Sci. 2022, 12, 9423.

[10] Isaac Ronald Ward, Ling Wang, Juan Lu, Mohammed Bennamounb, Girish Dwivedi and Frank M Sanfilippoa, "Explainable artificial intelligence for pharmacovigilance: What features are important when predicting adverse outcomes?", 2021 Elsevier.

[11] Avi Rosenfeld. 2021. Better Metrics for Evaluating Explainable Artificial Intelligence: Blue Sky Ideas Track. In Proc. of the 21th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021, IFAAMAS.

[12] Rohit Bharti, Aditya Khamparia, Mohammad Shabaz, Gaurav Dhiman, Sagar Pande, and Parneet Singh, "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning", Hindawi Computational Intelligence and Neuroscience Volume 2021, Article ID 8387680, https://doi.org/10.1155/2021/8387680.