Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2022) 30: 2054 – 2072

doi:10.55730/1300-0632.3924

A quantitative evaluation of explainable AI methods using the depth of decision tree

Nizar AHMED[1,*], **Adil ALPKOÇAK**[2]
[1]Department of Computer Engineering, Dokuz Eylül University, İzmir, Turkey
[2]Department of Computer Engineering, İzmir Bakırçay University, İzmir, Turkey

Received: 05.11.2021 • **Accepted/Published Online:** 15.06.2022 • **Final Version:** 28.09.2022

Abstract: It is necessary to develop an explainable model to clarify how and why a medical model makes a particular decision. Local posthoc explainable AI (XAI) techniques, such as SHAP and LIME, interpret classification system predictions by displaying the most important features and rules underlying any prediction locally. Therefore, in order to compare two or more XAI methods, they must first be evaluated qualitatively or quantitatively. This paper proposes quantitative XAI evaluation metrics that are not based on biased and subjective human judgment. On the other hand, it is dependent on the depth of the decision tree (DT) to automatically and effectively measure the complexity of XAI methods. Our study introduces a novel XAI strategy that measures the complexity of any XAI method by using a characteristic of another model as a proxy. The output of XAI methods, specifically feature importance scores from SHAP and LIME, is fed into the DT in our proposal. The DT will then draw a full tree based on the feature importance score decisions. As a result, we developed two main metrics that can be used to assess the DT's complexity and thus the associated XAI method: the total depth of the tree (TDT) and the average of the weighted class depth (ACD). The results show that SHAP outperforms LIME and is thus less complex. Furthermore, in terms of the number of documents and features, SHAP is more scalable. These results can indicate whether a specific XAI method is suitable for dealing with different document scales. Furthermore, they can demonstrate which features can be used to improve the performance of the black-box model, in this case, a feedforward neural network (FNN).

Key words: Explainable AI, medical multiclass classification, SHAP, LIME, decision tree, quantitative explainability evaluation

1. Introduction

Regarding any medical record, a doctor or physician is interested in knowing which medical entities or features are behind disease predictions or decisions. Hence, the need to create an automatic medical classification system which is more explainable and interpretable to human beings is obvious. This also means that any such medical system must be trustworthy and reliable in order to be useful in such sensitive medical decisions [1, 2]. Therefore, most NLP researchers are now attempting to develop automatic medical systems that are more explainable and understandable to human end-users.

Explainable AI (XAI) methods are divided into several categories based on the nature of the system they describe. Therefore, before employing any XAI method for our system, we must first ask ourselves some fundamental questions. For example, can the XAI method interpret a specific model or can it be generalized for use on any system? Alternatively, we must determine whether our XAI method explains the system's predictions

*Correspondence: nizar.ahmed@ceng.deu.edu.tr

locally or globally. More importantly, a decision must be made regarding which aspect of the model needs to be explained. Three categories define the type of XAI method based on that concept: premodel explanation, in-model explanation, and postmodel explanation (also known as posthoc models) [3–5]. Eventually, each explainable method must be quantitatively or qualitatively evaluated, which is a difficult but doable task.

An XAI system must be evaluated, especially when attempting to compare a newly created one to others in the literature. As a result, subjectively assessing the quality of the XAI system is dependent on the understanding of the end-user. The goal here is to satisfy the understanding of humans and to allow them to judge the effectiveness of your XAI method. As a result, there are two kinds of evaluation methods that rely on qualitative assessment. The first is known as a human-grounded assessment, and it necessitates direct interaction with the end-user, regardless of his or her knowledge of the system at hand. The second one is more practical, in that it selects more domain experts to judge the explainability of your XAI method. However, both techniques are expensive in time and effort and it may be possible to employ other practical and less time-consuming options, therefore, using quantitative evaluation methods is more practical and efficient. Functionally-grounded is an evaluation technique that does not need the interaction of humans to identify the quality of your XAI model [6]. It deals with the criteria of some measurements that serve as proxies, such as depth of the decision tree, model sparsity and uncertainty. Accordingly, quantitative techniques would be more automatic and practical than the previously mentioned qualitative techniques [7–10].

Aside from the significant time and effort required, one of the difficulties in qualitative XAI evaluation is that requesting human opinion to subjectively assess a particular explainability method can be biased to their desire and satisfaction. As a result, quantitative evaluation techniques during the evaluation process could be more objective and time saving. One of the main issues in the XAI research community is generalizing quantitative evaluation methods [11], because dealing with customized evaluation metrics cannot be applied to all other explainability methods. As a result, we require a mutual evaluation technique that is inclusively well suited to common explainability methods. Moreover, another issue is when different posthoc explanation approaches are increasingly being used to explain complicated models in high-stakes situations, it is essential to obtain a greater understanding of when and how the explanations produced by these methods conflict, and how such differences are addressed in practice. This problem was highlighted by recent study [12] as a XAI disagreement problem. Accordingly, Krishna et al. suggested developing a fundamental assessment criteria that can help practitioners quickly distinguish between a trustworthy and an unreliable explanation when there is a disagreement. Another important recent challenge in the XAI community is to measure the quality of explanations, i.e. causality [13], without the intervention of human judgment. Causality is the degree to which a human expert's explanation of a statement reaches a specific level of logical comprehension while maintaining effectiveness, efficiency, and satisfaction in each use case. Therefore, following this notion, Longo et al. [14] suggested a measurement scale that can be used to assess the quality of explanations objectively and automatically.

Following our discussion of the major issues and challenges with XAI evaluation techniques, we propose a novel and more practical method for quantitatively evaluating a specific XAI method. We chose to assess explainability of both: SHAP and LIME XAI methods because they are model agnostic and are extensively and frequently applied to a wide range of health-care data and industry domain to interpret deep learning models [15–17]. However, our main goal in this study is to create an automated evaluation metric that effectively and objectively measures XAI complexity. Moreover, this study does not attempt to prove that one XAI approach is superior to another; rather, it proposes that a feature of an existing ML model, such as DT, be used to select the

best XAI method for a given situation. As a result, we chose to measure XAI complexity using the depth of the decision tree because it is a practical and straightforward method [5]. Decision trees are one of the few machine learning algorithms that a human can easily track and analyze [18]. Hence, it will add simplicity and reliability when it will be used as an XAI evaluation measurement. There is a plenty number of XAI methods in literature but it is not easy to choose which one is proper for a given situation. Hence, we have to ask first, can the depth of DT be used to choose the right XAI methods for the given problem? This is our research hypothesis, which we will attempt to demonstrate in this study. Additionally, we intended to prove that our method can alleviate the previously discussed issues related to human-based evaluation techniques, generalization, causality and the disagreement problem.

Accordingly, our proposal begins with using the depth of the tree to evaluate an XAI's complexity by feeding its feature importance scores into the tree. As a result, the depth of the decision tree can indicate how complex the XAI method is [7, 8, 15]. To be more specific, we used a feedforward neural network (FNN) to classify a multiclass cardiovascular dataset (OHSUMED). Then, to explain our FNN's decisions, we used two cutting-edge explainability methods: SHapley Additive exPlanations Sheply (SHAP) and local interpretable model-agnostic explanations (LIME). The depth of the decision tree was then used to assess the complexity of the explanation methods.

The contributions of this study can be summarized as follows:

- To the best of our knowledge, this is the first idea that uses a feature of another model, such as a decision tree, as a proxy to evaluate the explainability method's complexity.
- Moreover, we performed a comparison between two XAI methods, SHAP and LIME, depending on the depth of their DT to measure XAI complexity.
- Additionally, the comparison of the two XAI methods was performed in the medical domain. More specifically, we worked with a multiclass classification related to a challenging cardiovascular diseases' dataset such as OHSUMED.
- Finally, as a byproduct intention of this study, XAI metrics, especially ACUDT metric, inspired us to use SHAP XAI importance scores as a dimensionality reduction mechanism to enhance the accuracy of our black box model. Accordingly, we only considered the top n features of OHSUMED dataset with the highest SHAP feature importance scores in the multiclass classification. As a result, the accuracy of our FFN model is remarkably increased from 50 to 92 percent.

The rest of the paper consists of the following: Section 2 provides information about the recent research regarding XAI and their evaluations. Section 3 explains in detail all the phases and steps of our method. Section 4 shows all the experiments and their settings. Section 5 presents all the results, findings, and discussions. And finally, Section 6 provides a summarization of our work and future work.

2. Related work

2.1. Explainable methods on ML predictions

Several articles dealt with explaining black-box models in a postoperation manner. Therefore, some researchers tried to explain the predictions of the black-box model using feature importance techniques. However, since feature importance scores are produced separately from the ML model training, they can be confusing or

misleading. As a result of this collision, numerous analytical techniques for calculating feature relevance and feature selection for ML models were emerged. Godin et al. [19], for example, used a feature importance method to investigate which patterns a neural network followed, after learning character-level features, to explain a word-level tagging and output of their neural network. On the other hand, Ge et al. [20] used a feature importance weighting approach in the medical domain. They distinguished which top 10 features are responsible for predicting intensive care unit (ICU) mortality. Moreover, Ribeiro et al. [3] created a feature importance technique that depends on the local model-agnostic method (LIME). LIME produces explanations that depend on estimating the model together with a "locally faithful" explainable representation. Additionally, it also uses the nature of some linear and explainable models, as proxy models, to interpret the predictions of a black-box model. Furthermore, LIME [3] employs local-based surrogate models that depend on input perturbation to produce explanations. To elaborate more, LIME takes an instance as an input and adds some noise to it so that it can produce more similar instances. After that, LIME uses linear models to learn from the new created instances to produce explanations. Furthermore, Blanco-Justicia et al. [21] proposed an explainable surrogate model that depends on microaggregation. Microaggregation derives explanations from the black-box model's decisions, while controlling their accuracy, and compares it with a decision tree as a surrogate model. Microaggregation is considered a type of decision tree but with limited depth. More specifically, they believe that this technique accomplishes a trade-off between the understandability of the original black-box model decisions and the representativeness of the surrogate model. Additionally, they think that their model will ensure the privacy of the instances used for training the black-box model.

2.2. Explainable AI in the medical domain

Some studies showed the importance of applying explainable methods to medical image analysis. Singh et al. [4] mentioned two main explainability groups that are used to explain deep neural networks in the medical images: attribution-based methods, and architecture or domain-specific methods. Attribution-based methods aim to discover which input features are affecting the target DNN neuron or the output neuron that is responsible for identifying the correct class in a classification problem. While domain-specific methods, sometimes called nonattribution-based methods, aim to develop an explainability method to a given problem rather than using preexisting attribution-based methods.On the other hand, Soares's paper [2] is a recent study that tries to explain the behavior of its classification model and to identify COVID-19 disease. Its dataset consists of CT scan images of patients infected with COVID-19. It used prototype-based learning in which a prototype, training images with a highly representative local peak of density and probability distribution, is used to explain the most important areas in the CT scans. The explainability of medical domain text classification also took a place in the literature. The main aim here is to extract and identify features from the medical text that contribute the most to define the black-box predictions' decisions [22]. Feng et al. [23] created an explainable clinical decision support text classifier. Their model consists of a CNN transformer that is responsible for extracting medical patient features from MIMIC III clinical records. Then, all the medical decisions are identified to be transformed later, by a two-layer transformer encoder, to recognize medical patterns among the features. Teng et al. [24] additionally built an interpretable ICD-9 classification system that aims to extract information from a knowledge graph. Their system consists of several parts: a multilayer CNN, a graph-based representation, an attention matching layer, and adversarial learning. Essentially, the knowledge graph has been constructed after applying some medical entity extraction techniques. This knowledge then is used in the attention layer to encode all the relevant medical features that contribute to identifying interdependency between ICD-9 codes from the MIMIC III dataset.

2.3. Evaluation of explainable methods

Quantitative explainability measurement has played a great part in evaluating explainability recently. Researchers have begun to design and perform some practical and quantitative evaluation metrics well suited to their explainability methods. Messalas et al. [25] tried to prove that the SHAP interpretability method is a faster alternative than LIME. Then they evaluated the performance of both techniques using six quantitative evaluation methods: runtime, consistency guarantees, identity, expressive power, translucency, and portability. All of those metrics depend on the features' correlations and importance. As a result, they prove that SHAP is better than LIME in time, consistency guarantees, and translucency, while having the same effect on the rest of the metrics. That means, time, consistency guarantees and translucency metrics are capable of detecting changes between SHAP and LIME, whilst the others are unable to do so. Similarly, Ramon et al. [26] made a comparison between three explainability methods: SHAP, LIME, and search for evidence counterfactuals (SEDC). SHAP uses a game theory method during the prediction process to consider the role of each specification or data instance. When the contribution is not equal, this strategy can equitably distribute the reward across all players (specifications). On the other hand, SEDC can explain any classification model's predictions in a counterfactual way. It focuses on finding the most relevant features during the explanation process. Ramon et al. [26] evaluated the effectiveness, which is represented by switching point and percentage explained metrics, and the efficiency, which is represented by the computational time, of the three explainable models. In the switching point, the amount of features that must be eliminated before the classification changes. It is the same as the counterfactual explanation's size. Alternatively, percentage explained is the fraction of positively predicted instances for which a counterfactual explanation smaller than 30 features is found. On the other hand, computation time is the time it takes for making the explanation in seconds. They found that SEDC consistently shows the best performance amongst them, it is fast and effective but not all the time, while SHAP and LIME recorded the second and third best performance, respectively. Lin [27] alternatively assessed the performance of three explainability methods: SHAP, LIME, and expected gradients (EG). They tested these methods on three classification problems: image, text, and speech recognition using deep CNN classifiers. After that, they measured the performance of the interpretable methods by assessing the impact score (IS) and and the impact coverage (IC). IS quantifies how well the critical factors, identified by an explainability method, reflect a given decision made by a network based on the impact over network decisions and confidences in the absence of these critical factors. Alternatively, IC quantifies the coverage of the detected essential variables on the negatively impacted factors. As a result, they found that the EG method recorded the best IS in the image and sound recognition classification problems, while LIME recorded the best score in the text classification problem. SHAP, on the other hand, obtained the best IC in the image classification problem, while EG and LIME got the best IC in text and sound recognition classification problems. Additionally, Antwarg et al. [28] compared SHAP and LIME after explaining the detection of anomalies by autoencoders. To assess the explainability methods, they depended on the mean reciprocal rank (MRR) measure to identify the noise position in the feature set. The lower the MRR, the better the explainability method is. As a result, SHAP recorded a better MRR score in comparison to LIME. Furthermore, El-Shawi et al. [29] compared six explainability methods after applying them to the predictions of a random forest black-box model. They utilized LIME, SHAP, Anchors, LORE, ILIME, and MAPLE XAI methods on healthcare tabular and textual datasets. They evaluated those methods using six quantitative assessment methods: similarity, time, trust, identity, stability, and separability. They concluded that SHAP performed the best in terms of time, similarity, and identity on the tabular dataset. Moreover, LIME recorded the best separability on the tabular and textual dataset and the

best similarity score in the textual dataset. Last but not least, Jesus et al. [30] performed a subjective (i.e. qualitative) assessment built on an application-based explainability evaluation metric. They constructed three basic questions and allowed domain experts to rate SHAP and LIME from one to five stars accordingly. These questions assessed how the XAI method could cover all relevant information on a decision, and whether XAI was useful in making that decision. Consequently, they found that people are more receptive to rate SHAP better than LIME in a fraud detection problem using the random forest as a black-box model.

As a result, XAI is still a new field and researchers are working on solving explainability problems especially on discovering new XAI evaluation techniques. However, the work on explainability in the medical field is still limited and needs more research to accomplish [31]. Additionally, making comparisons between the various XAI approaches is also important as it reveals which technique is more appropriate to utilize in a specific domain or dataset. Accordingly, we tried to tackle most of XAI's challenges and problems by creating a generic automated and objective XAI evaluation method. In addition, we provided a comparison between two benchmark XAI methods in a medical domain problem.

3. Methodology

The system in this research generally consists of three main phases. The chart in Figure 1 contains the structure that we followed to build our full system.

Phase 1: A classification and prediction phase that classifies the OHSUMED multiclass dataset into one of 23 cardiovascular diseases.

Phase 2: The prediction's explanation that interprets the output of the Blackbox model (i.e. the feedforward neural network, FNN). After that, the explanations are represented as features importance scores from one XAI tool like LIME or SHAP.

Phase 3: A quantitative evaluation method of explainability that depends on the depth of the decision tree.

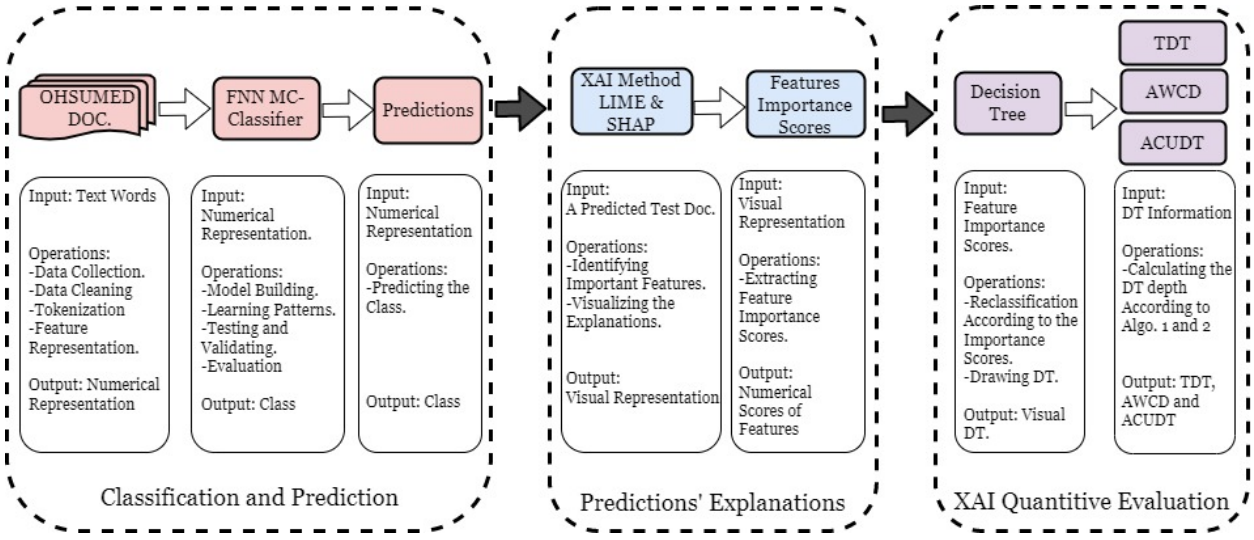

Figure 1. General system architecture.

3.1. Classification and prediction system

3.1.1. Dataset description

We employed the OHSUMED dataset in this study. OHSUMED is a collection of medical abstracts from the MEDLINE database that provides information on 23 cardiovascular diseases [32, 33]. Additionally, each document is an unstructured sequence of words. There are 6900 documents in total. Furthermore, our dataset has a balanced distribution of documents per class. There is a minimum of 232 documents per class, class C23, max 368 documents per class, class C22, a variance of 1204.79, and a standard deviation of 34.71009. We split the data into 80% training and 20% testing for the classification task. Table 1 provides a detailed description of the dataset.

Table 1. Basic statistics about OHSUMED dataset.

Descriptions	**OHSUMED dataset**
Total number of documents	6900
Total number of sentences	57,647
Total number of tokens	1,244,089
Total number of characters	8,097,928
Max document size in sentence	23
Max document size in tokens	605
Max document size in characters	4025
Average document size in sentence	5
Average document size in token	180
Average document size in characters	1173
Total number of classes	23

3.1.2. Multiclass classification system

Our multiclass classification model used in this study is a simple feedforward neural network with one input layer, one hidden layer, and one output layer. Moreover, we established several hyperparameters that provide us with high-accuracy and low-loss scores:

- Activation function in the input and hidden layers: Relu.
- Activation function in the output layer: Softmax.
- Loss function: Sparse categorical crossentropy.
- Optimizer: Adam.
- Epochs = 15.
- Batch size =128.

3.2. Explainability methods

We worked on explaining the black-box model's output with multiclass classification predictions. As a result, we attempted to use local explainability methods to explain the FNN's predictions and decisions. Local

explainability techniques such as LIME and SHAP are examples of local explainability methods. Algorithm 1 demonstrates how we used the LIME and SHAP XAI methods to obtain the explanations as feature importance scores for each document in OHSUMED.

3.2.1. Feature importance XAI method: SHAP and LIME

In general, feature importance is defined as the degree to which features are related to the black-box model's prediction decisions. Intuitively, the more the model's decisions are based on the feature, the more sensitive its predictions become to changes in that feature. We can assess the significance of a feature by first calculating the increase in error after perturbing the feature. Second, a feature is considered important if the model's error increases proportionally to the perturbing value of that feature. That is, the model's predictions are extremely sensitive to that feature, and thus it is deemed "important" [7].

Local interpretable model-agnostic explanations (LIME) is one of the most famous feature importance techniques that is used in the literature [3, 7]. LIME uses an explainable model as a surrogate (usually, a model that has interpretable nature such as linear models, trees, or rule-based models) to locally explain a black-box model. Its work principle starts with selecting an instance of interest, a document, whose explanation is desired after obtaining the black-box prediction. Secondly, LIME perturbs or adds noise to the dataset and obtains the black-box predictions on the disturbed data. Then, LIME weighs new data samples by calculating their closeness to the instance of interest. After that, it fits a weighted interpretable model, the surrogate model, on the perturbed dataset. And finally, LIME explains the black-box predictions by interpreting the local model.

SHapley Additive exPlanations (SHAP), on the other hand, connects game theory principles with local explanations to represent the only possible consistent local features attributed with the black-box decisions [7, 34]. The predictions of the black-box model can be explained by assuming that each feature is a "player" in a game, and each prediction is considered as the "payout" of that game. The Shapley value, which is a big factor in the game theory perspective, represents how to fairly assign the "payout", or predictions, to each "player" or feature. SHAP's working principles start with representing a prediction problem of a single data instance as a "game". Then, the "gain" or "payout" of the game is the actual instance predictions subtracted by the average predictions of all the other instances. And finally, the "players" are the features that were collaborated together to receive the "gain".

Algorithm 1 The calculation of LIME and SHAP XAI importance scores.

Input: OHSUMED documents D = (d_1, d_2, d_3 ..... d_n) each related to one class c_i where C = (c_1, c_2, c_3 .... c_n).

Output: A matrix X contains: a set of documents d_n represented by its features f_m and its related LIME/SHAP importance scores s_m. X=(s_1, s_2, s_3 ....., s_m)

```
for d_i in D do
    for f_m in d_i do

        X ← Calculate LIME/SHAP importance score s_m.

    end for
end for
```

3.3. A quantitative evaluation technique of explainability

Each feature in the previous phase is associated with a single importance score from a specific explainability method. As a result, we employ the depth of the decision tree (DT) as a metric for measuring the complexity of the classification system. Instead of using the original sequences of OHSUMED documents, we use the explainability model's feature importance scores as an input to the decision tree. Then we let the DT classify the OHSUMED documents based on the importance of each feature. Following that, we use the resulting rules and graph to determine the depth of the tree and thus whether or not the explainability technique is complex. Algorithm 2 shows how we calculate each XAI evaluation metric.

For example, the tree in Figure 2 is the result of feeding XAI feature importance scores into a decision tree-based classifier. The root of the tree contains the entire set of data used in the classification problem, which in this case is four samples. Following that, each branch stores information about each node, such as the number of samples in each node and the Gini score. Additionally, this tree has four leaves, each of which represents a class or the target of a specific decision. In our case, the four classes are C13, C06, C19, and C22. As a result, this tree can hold three different XAI evaluation measurements, two for XAI complexity and one for DT performance. Thus, the total depth of the decision tree is the first complexity measurement (TDT). The second is the average of the weighted class depth (AWCD). Finally, the accuracy of the DT classification model (ACUDT) is used to evaluate the decision tree's performance.

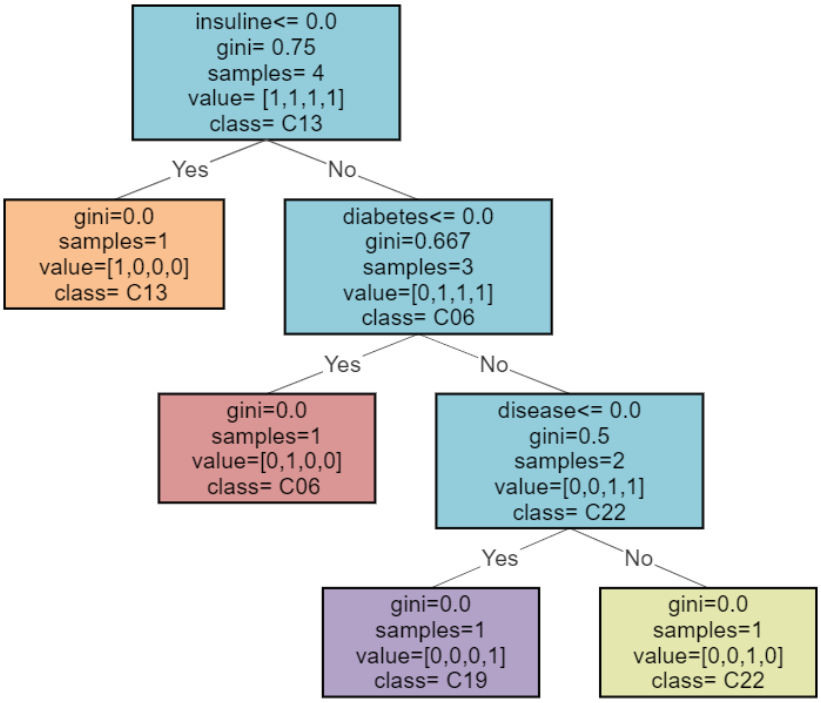

Figure 2. A decision tree graph of four document samples.

3.3.1. The total depth of the DT

Some researchers have suggested that the total depth of a decision tree can be used as a complexity indicator [7, 8, 15]. The model becomes more complex as the depth of the DT increases. Hence, the total depth of the

tree in Figure 2 is:

$$TDT = Count(L), \tag{1}$$

where L represents each level of the tree; the root is not included. Consequently, the total number of tree levels TDT in Figure 2 tree is three.

3.3.2. The average of the weighted class depth

Instead of the total depth of the tree, this measurement can tell us about the complexity in a class-based depth. As previously stated, this study comprises 23 cardiovascular classes. In the resulting decision tree, each class represents a leaf. Therefore, each leaf can be found at a specific level. Class C13, for example, exists in the DT's first level. In level two, class C06 is a leaf. C19 and C22 are in the tree's level three. In this case, we compute the average of the weighted (by number of samples) class depth (AWCD) as follows:

$$AWCD = \frac{1}{N} \sum_{D=1}^{d} L_{class} \times C_S, \tag{2}$$

where N is the total number of data samples, in this case four in Figure 2. D is the total number of leaves in the tree; the tree uses only four classes in total. Each class leaf's level is represented by Lclass. Hence, each class leaf, C13, C06, C19, and C22 is in the first, second, and third levels, respectively. Finally, the number of data samples in the current leaf is denoted by CS. In that case, the number of samples in each class leaf, C13, C06, C19, and C22, is equal to one. As a result, Equation 2 can be mathematically calculated as follows:

$AWCD = \frac{1}{N} \times [[L_{C13} \times Saples_{C13}] + [L_{C06} \times Saples_{C06}] + [L_{C19} \times Saples_{C19}] + [L_{C22} \times Saples_{C22}]]$

$AWCD = \frac{1}{4} \times [[1 \times 1] + [2 \times 1] + [3 \times 1] + [3 \times 1]]$

$AWCD = \frac{9}{4} = 2.25.$

3.3.3. The accuracy of DT

We can also consider the DT classifier's accuracy (ACUDT) to represent how the DT performs when XAI's feature importance scores are used as inputs to the DT. In that case, we can rely on the DT's ability to correctly classify the documents. Furthermore, this metric can be used to determine which features will improve the performance of the black-box classifier.

4. Experimental setup

In the methodology section, we provided all the information on our FNN's hyperparameters. In addition, we used Python 3.6 and the spider platform to write our code. You can find the full code of the system in our GitHub repository https://github.com/nizar3112/DT_XAI_EVALUATION.git. We performed several experiments to investigate the following:

1. Based on explainability feature importance scores, how DT accuracy scales with the number of documents.
2. How the number of top feature sets with the highest importance score affects the DT accuracy.
3. Whether the number of features and/or documents affect the total depth of the DT.
4. How the number of features and/or documents directly impacts the average of the weighted class depth of the DT.

Algorithm 2 The calculation of each XAI evaluation metric.

Input: The matrix X, which is resulted from Algorithm 1, where each raw is attached with its related class c_i.
Output: The three XAI evaluation metrics.

Apply a Decision Tree (DT) algorithm to the matrix X and its related class c_i.
Draw DT that consists of multiple levels L= (l_1, l_2, l_3...,l_z), and its leaves as c_i.

TDT ← Count (L).

for each class c_i and its related number of data samples $samp_s$ in a particular DT level L c_i **do**

    AWCD ← $\frac{1}{X}$ $\sum_{L}^{1}$ L c $_i$ × $samp_s$

end for

ACUDT ← DT accuracy score.

Algorithm 3 The calculation of LIME and SHAP XAI importance scores.

Input: OHSUMED documents D = (d_1, d_2, d_3.....d_n) each related to one class c_i where C = (c_1, c_2, c_3....c_n).
Output: A matrix X contains: a set of documents d_n represented by its features f_m and its related LIME/SHAP importance scores s_m. X=(s_1, s_2, s_3.....s_m)

for d_i in D **do**
    for f_m in d_i **do**

        X←Calculate LIME/SHAP importance score s_m.

    end for
end for

4.1. LIME and SHAP experiments

To obtain the explanations for each document, we used the LIME and SHAP libraries, which are compatible with our FNN infrastructure. As a result, each document is now represented by a series of feature importance scores derived from the LIME and SHAP XAI methods. However, we set up several experiments in the following manner:

1. Document-based experiments: For each XAI method, we set up four experiments based on the number of documents. We assigned 100 documents (baseline), 500 documents, 1400 documents (only test documents), and the entire set of 6900 documents (train and test sets). This parameter, on the other hand, can show us how the accuracy and depth of DT are affected on a document basis.

2. Experiments with the top feature set: We began to audit the number of top features with the highest importance scores as a dynamic parameter. This parameter can help us understand how to adjust the accuracy and DT depth. Accordingly, we set that parameter to 20, 50, 100, and 500, the maximum number of features that a document can have.

3. Experiments to enhance the accuracy of the multiclass classification model using SHAP as a dimensionality reduction method: For each document in OHSUMED dataset, we choose the top 100, 200, 300, and 400

features after calculating their SHAP importance scores in the multiclass classification. This step was responsible for increasing the accuracy of our FNN especially when the number of FNN epochs increased to 30.

5. Results and discussion

This section summarizes our findings in terms of two different factors: document and feature distribution-based parameters. As previously stated, the change in the number of documents can be used as an indicator of XAI complexity, while the change in the number of features can be used as a performance metric to indicate whether the classifier is performing well or poorly.

5.1. Document scalability effect on the evaluation metrics

We show the effect of increasing the number of documents, primarily from 100 to the entire 6900 sets of data. Figure 3 depicts how the DT's total depth, average weighted class-based depth and accuracy vary with document size. We discovered that the average weighted class-based depth, total depth, and accuracy are all proportional to the number of documents. For both SHAP and LIME, the three metrics scores increase and decrease as the number of documents increases or decreases, respectively. Furthermore, we found that SHAP outperforms LIME in terms of DT complexity (as measured by the average weighted class depth and total DT depth) and performance (as measured by DT accuracy). Furthermore, there is a significant difference in DT accuracy between SHAP and LIME However, both SAHP and LIME show a slight difference in terms of the average weighted class-based depth and DT total depth, although the former performed better.

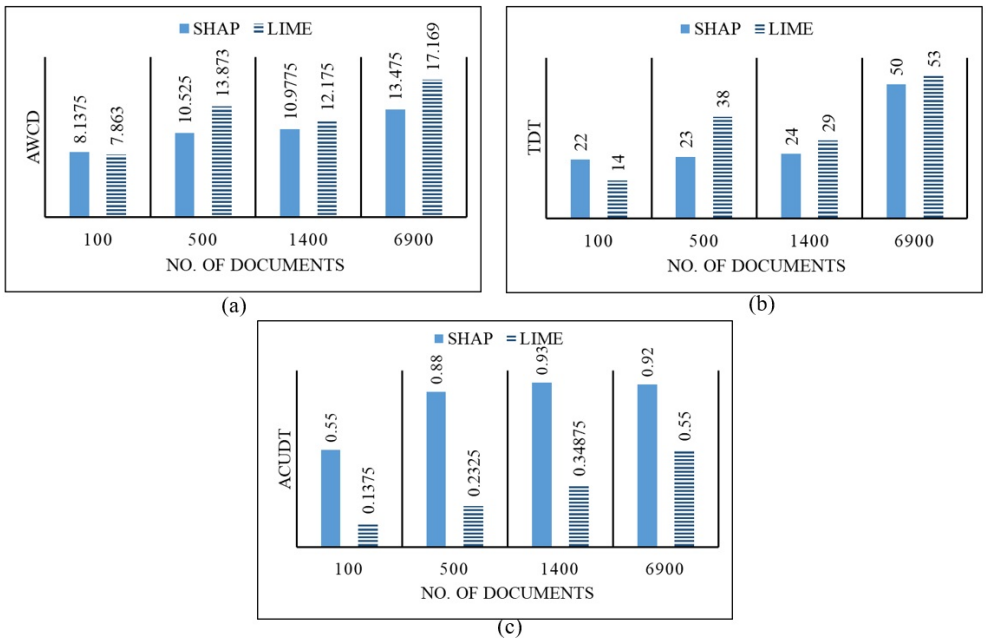

Figure 3. The results of (a) average weighted class depth, AWCD; (b) total DT depth, TDT; and (c) DT accuracy, ACUDT, with each set of documents.

5.2. Feature distribution effect on the evaluation metrics

Figure 4 illustrates how the number of features played an important role in the evaluation process. We conclude that both complexity metrics, average weighted class depth and DT total depth, produce the highest scores with

SHAP. As a result, SHAP receives the lowest scores for both depth metrics and is thus less complex than LIME. The complexity of both XAI methods, on the other hand, is inversely proportional to the number of features. In other words, as the number of features decreases, the complexity of XA increases, and vice versa. On the other hand, increasing the number of features increases DT accuracy and vice versa. We also found that the differences in complexity metrics between SHAP and LIME is not significant, with only a three-level difference when 500 features are considered. While the DT accuracy differed noticeably between the two XAI methods, SHAP achieves the highest DT accuracy, 92 percent, using 500 and 200 features, whereas LIME achieves 55 percent and 54 percent using the same number of feature sets. Furthermore, as the number of top features increases, the tree transforms from a wide and sparse tree with up to 500 features, to a deeper one with only 20 features.

Generally, Figure 5 shows the big picture of the three evaluation metrics on SHAP and LIME. It illustrates that SHAP, in general, outperforms LIME in terms of the three measurements.

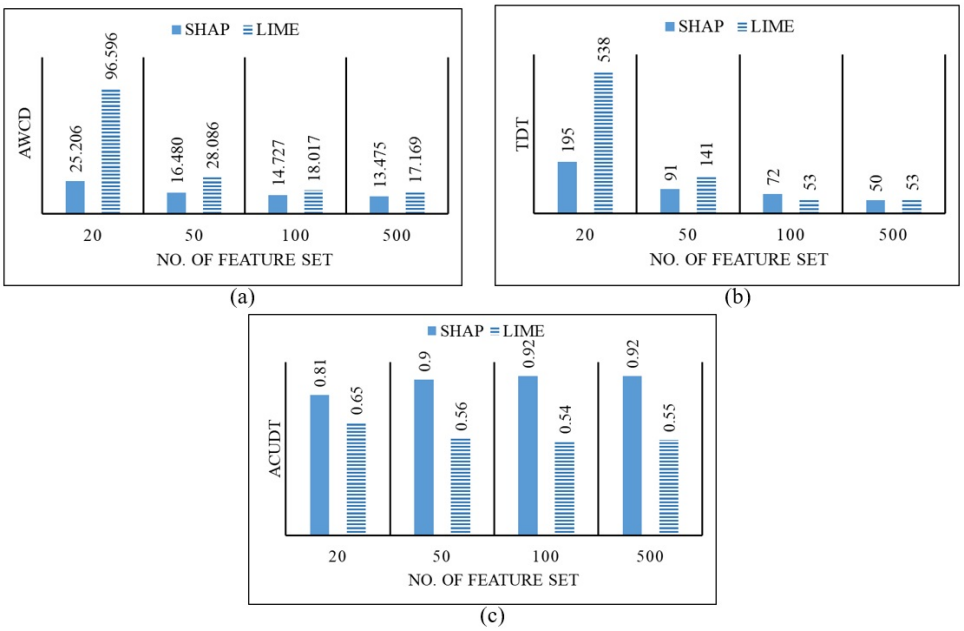

Figure 4. The results of (a) average class depth; (b) total DT depth; and (c) DT accuracy, with each set of features.

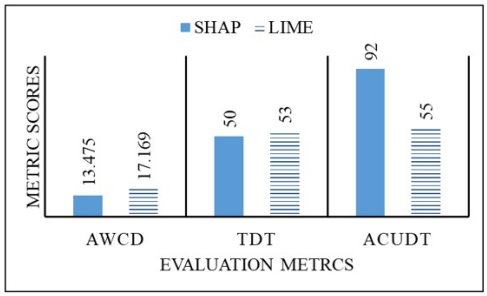

Figure 5. A comparison of SHAP and LIM for each evaluation metric.

5.3. Using SHAP as a dimensionality reduction

When it comes to classifying the entire set of OHSUMED data, our FNN model achieves only 49.8 percent accuracy. It means that our black-box model is less effective when all features are taken into account. However,

for the sake of XAI, this study is concentrating on evaluating XAI methods and comparing them in terms of quantitative evaluation measurements. As a result, the goal of this study is not to evaluate the performance of the black-box classifier, but rather to deal directly with the XAI methods and their complexity. Although, it is not intended to improve the accuracy of our multiclass classifier, and yet as a byproduct goal, we got the inspiration from the ACUDT XAI metric to increase the efficiency of our black-box model. As it is elaborated in experiment no. 3, we tried to enhance the accuracy of FNN multiclass classification by employing SHAP as a dimensionality reduction mechanism. Therefore, the black-box accuracy is reaching its best, i.e. recording 92.1 percent accuracy, when the number of features are reduced to 200 and the number of epochs is increased to 30. Table 2 shows all the results of using different number of feature set with increasing the number of FNN epochs from 15 to 30.

5.4. Discussion

Based on the previous results, we concluded that SHAP outperforms LIME throughout all three evaluation metrics. The nature of calculating feature importance scores differs between the two XAI methods. These distinctions could explain the main differences between the two methods. Furthermore, as we discussed in subsection 2.3, most of the literature research potentially supports SHAP outperforming LIME in other quantitative and qualitative evaluation metrics [25, 28, 30].

Moreover, the total depth of the decision tree and the average weighted class depth are good indicators of the complexity of the XAI method. Figure 5 shows that SHAP has a more stable and lower total tree depth than LIME. However, the AWCD metric is more specific than TDT in determining the degree of complexity. Table 3 shows, for example, how the TDT metric is stable and does not change; TDT equals 22 levels deep because different sets of features were used. While TDT is stable, AWCD can show differences in class depths based on each leaf depth, because the depth of each class can be found at a different level, and hence the average of the depth's weighted sum can be more useful in defining complexity measurement.

Table 2. The accuracy score of the FNN after selecting the features with top SHAP importance scores.

No. of features	**Accuracy score – 10-fold crossvalidation**		
	15 epochs	**20 epochs**	**30 epochs**
100	62.42	79.71	91.48
200	64.83	79.56	92.1
300	56.36	77.02	91.63
400	56.02	76.4	91.13

Table 3. TDT and AWCD evaluation metrics on 100 documents and multiple sets of features.

No. of top features	**20**	**50**	**100**	**500**
TDT metric	22	22	22	22
AWCD metric	9.75	8.61	8.51	**8.14**

Furthermore, document scalability is an indicator that reflects how XAI methods work with different numbers of documents. SHAP is more scalable and can handle documents of more varied sizes than LIME. On the other hand, feature distribution can help us learn more about the effects of important scores from SHAP and LIME on the decision tree's classification accuracy. This factor for different feature sets can demonstrate

how SHAP, with its importance scores, is more efficient than LIME in positively affecting the decision tree's accuracy.

Figure 5 shows that the accuracy of the DT with the SHAP method is 92 percent on the full set of documents and 500 feature sizes, while LIME is at 55 percent. We also believe that this accuracy is an XAI accuracy and that it cannot be compared to the classification accuracy of the black-box model, i.e. the FNN network. Furthermore, XAI importance scores are generated based on how a specific black-box model responds to the entered data. Accordingly, we cannot claim that the decision tree is more robust and powerful than FNN in classifying OHSUMED documents because both accuracies serve different purposes and need different model configurations. Alternatively, the decision tree's XAI accuracy, i.e. the ACUDT metric, can be utilized as a guide to help us choose only the most significant features as an input to the black-box model [35]. In our situation, the ACUDT metric inspired us to improve our FNN's accuracy by controlling and selecting the most significant features during multiclass classification. Furthermore, there are a number of irrelevant and noisy features that are accountable for degrading the FNN accuracy. Hence, we controlled this problem by using SHAP as a feature selection technique to enhance the performance of the multiclass classification model. Accordingly, the resulted features forced the classification model to be more discriminative so that it can identify the pattern of each class more appropriately.

For comparison, Tables 4 and 5 provide a comparison between our XAI evaluation method, i.e. DT depth method, and other some recent papers in the literature. Although there are some differences between our method and theirs, such as the use of different datasets and black-box models, yet they are comparing among several XAI methods as we do. The main aim of the comparison in Table 4 is to show how the other recent research drew the same conclusion as we did, which is that SHAP outperforms other XAI methods. Although some papers such as Ramon et al. [26] indicate that SEDC XAI method is the superior XAI method, they also proved in their research that SHAP was better than LIME as we did in our method. That is, on comparison, the effectiveness of SHAP tool is validated by not only the XAI evaluation metrics implemented in the literature but also by our evaluation method.

Table 4. A comparison of our method and other different XAI evaluation systems in the literature.

Paper	**Dataset type**	**XAI methods**	**Evaluation methods**	**The best XAI method**
Our method	Text	LIME and SHAP	DT depth	SHAP
[25]	Text	LIME and model-agnostic Shapley value explanations (MASHAP)	Run time, consistency guarantees, identity, expressive power, and translucency	MASHAP
[26]	Text	Counterfactual LIME-C, counterfactual SHAP-C and search for evidence counterfactuals (SEDC)	Switching point	SEDC
[28]	Text	LIME and SHAP	Mean reciprocal rank (MRR) Measure	SHAP
[29]	Tabular and text	LIME, SHAP, Anchors, LORE, ILIME and MAPLE	Similarity and bias detection	Tabular: SHAP and text: MAPLE
[30]	Text	LIME and SHAP	Application-grounded	SHAP

Table 5. A comparison of the weakness and strength of our method and other XAI evaluation metrics in literature.

Evaluation metric	**Weekness**	**Strength**
Depth of DT (our proposal)	Tree may overfit depending on the number of data samples	Automatic, objective, ML based, visual dependent and easy to track approach
Run time	Dependent on the no. of explanations	Automatic and more common for measuring performance
Consistency guarantees	Manual and subjective	Feature dependent approach
Identity	Manual and subjective	Groups the data points according to the same explanation
Translucency	Manual and limited to some XAI methods	Unique in terms of showing the inner working of the model
Expressive power	Manual and subjective	It shows the effect of the explanation language and visualization.
Switching point	Manual and subjective	It shows the method's ability to rank features from high to low relative importance.
Mean Reciprocal Rank (MRR)	Manual selection of noise features	Objective measure
Similarity	Manual	Objective measure
Bias detection		
Application-grounded	Manual and subjective	Human-in-the-loop

Besides the overfitting problem as a side effect of increasing the size of the data samples, our XAI metric approach provides multiple advantages compared to the other XAI assessment metrics in the literature. As it is illustrated in Table 5, our approach is a quantitative, visual, and machine learning-based method that efficiently evaluates the XAI complexity based on the depth of the DT. However, other XAI evaluation metrics, such as mean reciprocal rank (MRR) [28], similarity, and bias detection [29], can objectively measure explainability. Others, such as consistency guarantees [25] and switching point [26], are feature-based approaches that effectively analyze XAI complexity by altering, adding, deleting, and rating essential data features. Furthermore, the identity metric [25] benefits by grouping data points based on similar explanations. Furthermore, translucency [25] measures the degree to which the XAI can explore the inner working of the black-box model.

On the other hand, the following XAI measurements in the literature have the drawbacks of being manual and subjective: consistency guarantees, identity, expressive power [25], switching point, and application-grounded [30]. These measurements are nonautomatic due to the manual selection of relevant features. In addition, the measurements are subjective because they depend on human judgment. Other XAI evaluation metrics, such as mean reciprocal rank (MRR) [28], similarity, and bias detection [29], are calculated manually but still have a quantitative nature. As a result, our XAI evaluation metric considers all the drawbacks mentioned above and adds additional components to make the measure more straightforward and robust.

Finally, researchers may debate about whether XAI representation, such as XAI visualization, should be included in the XAI evaluation process. However, this can sway human judgment and bias it toward a particular XAI method. Nevertheless, in our opinion, the most important part of the explanations are the features and their role in driving a specific prediction decision. As a result, the key issue behind any black-box explanation is how the XAI methods can define the most important features. According to Jesus et al. [30], "another important consideration is the biases that may emerge if explanation methods are distinguishable due to some factor (e.g., their representation). Preventing their representational differences is thus a necessary precaution toward isolating the quality and relevance of the explanation methods from all other potential visual factors". Hence, using quantitative XAI evaluation measures will be more robust in providing an interpretability assessment that is not influenced by human sentiments.

6. Conclusion

Explainability AI (XAI) methods have the capability of interpreting black-box model's prediction decisions. Some XAI strategies focused on feature importance scores to assign significance to features based on how responsive the black-box model is to that feature. As a result, some local and global XAI techniques rely entirely on that mechanism. Then, evaluating XAI methods becomes necessary, which can be done using either subjective or objective methods. Subjective or qualitative evaluation methods rely on human judgment and can be time-consuming and labor-intensive, while objective or quantitative methods automatically evaluate explainability to save time and effort.

This paper proposes a quantitative XAI evaluation technique for measuring the complexity of the local-based XAI method, which is heavily reliant on feature importance scores. The depth of the decision tree (DT) is used as a complexity metric to assess and compare the complexities of two XAI methods: SHAP and LIME. The more complex the tree, the deeper it is. The explanations, or importance scores, of the XAI method, are fed into the decision tree. Following that, the tree will draw decisions based on the input explanations, allowing us to calculate its depth to measure complexity. Consequently, we developed two complexity metrics: one based on the total depth of the DT (TDT), and the other on the average weighted depth of each DT's leaf or class (AWCD). On the other hand, there is another metric that is based on the DT's accuracy (ACUDT) which can be used to determine the appropriate number of features to increase the performance of the DT classifier. The DT accuracy can help us figure out how to control the black-box accuracy's efficiency in the future.

The results show that the DT of the SHAP XAI method is less deep and less complex than that of the LIME method. Furthermore, SHAP outperforms LIME in terms of document scalability and feature sensitivity. To ensure that we are on the right track, our findings agree with several studies that demonstrate SHAP's effectiveness, particularly when compared to LIME. We intend to compare some global XAI methods using our quantitative metric.

References

[1] Kishlay J, Yaqing W, Guangxu X, Aidong Z. Interpretable word embeddings for medical domain. Proceedings - IEEE International Conference on Data Mining. ICDM 2018; 1061-1066. doi: 10.1109/ICDM.2018.00135

[2] Angelov P, Soares E. Explainable-by-design approach for covid-19 classification via ct-scan. MedRxiv 2020; 2020-04. doi: 10.1101/2020.04.24.20078584

[3] Ribeiro M, Singh S, Guestrin C. Model-agnostic interpretability of machine learning. ArXiv 2016; 1606.05386.

[4] Singh A, Sengupta S, Lakshminarayanan V. Explainable deep learning models in medical image analysis. Journal of Imaging 2020; 6 (6):52.

[5] Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 2020; 58:82-115. doi:10.1016/j.inffus.2019.12.012

[6] Islam S, Eberle W, Ghafoor S. Towards quantification of explainability in explainable artificial intelligence methods. Proceedings of the 33rd International Florida Artificial Intelligence Research Society Conference, FLAIRS 2020; 75-81. doi:10.1016/j.inffus.2021.05.009

[7] Molnar C. Interpretable machine learning. Leanpub. Lulu.com 2020.

[8] Carvalho D, Pereira E, Cardoso J. Machine learning interpretability: A survey on methods and metrics. Electronics 2019; 8 (8):1-34. dio:10.3390/electronics8080832

[9] Mohseni S, Block J, Ragan E. Quantitative evaluation of machine learning explanations: A human-grounded benchmark. In 26th International Conference on Intelligent User Interfaces 2021; 22-31. doi:10.1145/3397481.3450689

[10] Vander J, Nieuwburg E, Cremers A, Neerincx M. Evaluating XAI: A comparison of rule-based and example-based explanations. Artificial Intelligence 2021; 291:103404. doi:10.1016/j.artint.2020.103404

[11] Doshi-Velez F, Kim B. Considerations for evaluation and generalization in interpretable machine learning. In Explainable and interpretable models in computer vision and machine learning. Springer 2018; 3-17. doi:10.1007/978-3-319-98131-4-1

[12] Satyapriya K, Tessa H, Alex G, Javin P, Shahin J and et al.The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective. ArXiv preprint 2022; 2202.01602.

[13] Andreas H, Randy G, Freddy L, Peter K,and Andreas H. Measuring the Quality of Explanations: The System Causability Scale (SCS) .KI - Künstliche Intelligenz 2020; 34:193-198. doi:10.1007/s13218-020-00636-z

[14] Luca L, André C,and Heimo M. Explainable artificial intelligence: Concepts, applications, research challenges and visions.International Cross-Domain Conference for Machine Learning and Knowledge Extraction 2020; 1-16. doi:10.1007/978-3-030-57321-8-1

[15] Zhou J, Gandomi A, Chen F, Holzinger A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics 2021;10 (5):1-19. doi:10.3390/electronics10050593.

[16] Hailemariam Y, Yazdinejad A, Parizi R, Srivastava G, Dehghantanha A. An Empirical Evaluation of AI Deep Explainable Tools. In 2020 IEEE Globecom Workshops 2020; 1-6. doi:10.1109/GCWkshps50303.2020.9367541.

[17] Ghassemi M, Oakden-Rayner L, Beam A. The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health 2021; 3 (11):745-50. doi:10.1016/s2589-7500(21)00208-9.

[18] Hearty A, Gibney M. Analysis of meal patterns with the use of supervised data mining techniques—artificial neural networks and decision trees. The American journal of clinical nutrition 2008; 88 (6):1632-42. doi:10.3945/ajcn.2008.26619

[19] Godin F, Demuynck K, Dambre J, De Neve W, Demeester T. Explaining character-aware neural networks for word-level prediction: Do they discover linguistic rules?. ArXiv preprint 2018; 1808-09551. doi:10.18653/v1/d18-1365

[20] Ge W, Huh J, Park Y, Lee J, Kim Y et al. An Interpretable ICU Mortality Prediction Model Based on Logistic Regression and Recurrent Neural Networks with LSTM units. In American Medical Informatics Association AMIA Annual Symposium Proceedings 2018; 460. doi:10.3390/jimaging6060052

[21] Blanco-Justicia A, Domingo-Ferrer J, Martinez S, Sanchez D. Machine learning explainability via microaggregation and shallow decision trees. Knowledge-Based Systems, Elsevier B.V 2020; 194:105532. doi:10.1016/j.knosys.2020.105532

[22] Singh D, Kumar V, Yadav V, Kaur M. Deep neural network-based screening model for COVID-19-infected patients using chest X-ray images. International Journal of Pattern Recognition and Artificial Intelligence 2021; 35 (03):2151004. doi:10.1142/s0218001421510046

[23] Feng J, Shaib C, Rudzicz F. Explainable clinical decision support from text. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) 2020; 1478-1489. doi:10.18653/v1/2020.emnlp-main.115

[24] Teng F, Yang W, Chen L, Huang L, Xu Q. Explainable prediction of medical codes with knowledge graphs. Frontiers in Bioengineering and Biotechnology 2020; 8:1-11. doi:10.3389/fbioe.2020.00867

[25] Messalas A, Aridas C, Kanellopoulos Y. Evaluating MASHAP as a faster alternative to LIME for model-agnostic machine learning interpretability. In 2020 IEEE International Conference on Big Data (Big Data) 2020; 5777-5779 doi:10.1109/BigData50022.2020.9378034

[26] Ramon Y, Martens D, Provost F, Evgeniou T. A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: SEDC, LIME-C and SHAP-C. Advances in Data Analysis and Classification, Springer, 2020; 14 (4):801-19. doi:10.1007/s11634-020-00418-3

[27] Lin Z, Shafiee M, Bochkarev S, Jules M, Wang X et al. Do explanations reflect decisions? A machine-centric strategy to quantify the performance of explainability algorithms. ArXiv preprint 2019; 1910-07387.

[28] Antwarg L, Miller R, Shapira B, Rokach L. Explaining anomalies detected by autoencoders using SHAP. ArXiv preprint 2019; 1903.02407.

[29] ElShawi R, Sherif Y, Al-Mallah M, Sakr S. Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. Computational Intelligence 2021; 37 (4):1633-50. doi:10.1111/coin.12410

[30] Jesus S, Belém C, Balayan V, Bento J, Saleiro P et al. How can I choose an explainer? An Application-grounded Evaluation of Post-hoc Explanations. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 2021; 805-815. doi:10.1145/3442188.3445941

[31] Lim T, Tay K, Huong A, Lim X. Breast cancer diagnosis system using hybrid support vector machine-artificial neural network. International Journal of Electrical and Computer Engineering 2021; 11 (4):2088-8708.

[32] Hersh W, Buckley C, Leone T, Hickam D. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In SIGIR'94, Springer 1994; 192-201. doi:10.1007/978-1-4471-2099-5-20

[33] Ahmed N, Dilmaç F, Alpkocak A. Classification of Biomedical Texts for Cardiovascular Diseases with Deep Neural Network Using a Weighted Feature Representation Method. In Healthcare 2020; 8 (4):392. doi:10.3390/healthcare8040392

[34] Lundberg S, Lee S. A unified approach to interpreting model predictions. In Proceedings of the 31st international conference on neural information processing systems 2017; 4768-4777.

[35] Marcílio W, Eler D. From explanations to feature selection: assessing SHAP values as feature selection mechanism. In 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), IEEE 2020; 340-347. doi:10.1109/SIBGRAPI51738.2020.00053