IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE SYSTEMS, MAN AND CYBERNETICS SOCIETY SECTION

Received 28 November 2023, accepted 18 December 2023, date of publication 28 December 2023,
date of current version 24 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3347796

RESEARCH ARTICLE

Toward Building Trust in Machine Learning Models: Quantifying the Explainability by SHAP and References to Human Strategy

ZHAOPENG LI[1], MONDHER BOUAZIZI[2], (Member, IEEE), TOMOAKI OHTSUKI[2], (Senior Member, IEEE), MASAKUNI ISHII[1,3], AND ERI NAKAHARA[4]

[1]Graduate School of Science and Technology, Keio University, Yokohama 223-8522, Japan
[2]Faculty of Science and Technology, Keio University, Yokohama 223-8522, Japan
[3]Social Informatics Laboratories, Nippon Telegraph and Telephone Corporation, Tokyo 239-0847, Japan
[4]NTT Smart Data Science Center, Nippon Telegraph and Telephone Corporation, Tokyo 108-0075, Japan

Corresponding author: Zhaopeng Li (zhaopeng_lee@keio.jp)

ABSTRACT Local model-agnostic Explainable Artificial Intelligence (XAI), such as LIME or SHAP, has recently gained popularity among researchers and data scientists for explaining black box Machine Learning (ML) models. In the industry, practitioners focus not only on how these explanations can validate their models but also on how they can help maintain trust from end-users. Some studies attempted to measure this ability by quantifying what they refer to as the explainability or interpretability of ML models. In this paper, we introduce a new method for measuring explainability with reference to an approximated human model. We develop a human-friendly interface to strategically collect human decision-making and translate it into a set of logical rules and intuitions, or simply annotations. These annotations are then compared with the local explanations derived from common XAI tools. Through a human survey, we demonstrate that it is possible to quantify human intuition and empirically compare it to a given explanation, enabling a practical quantification of explainability. By relying on this new method, we identified several potential flaws in today's ML selection process. Furthermore, we demonstrate how our method can help to better evaluate ML models.

INDEX TERMS Explainable artificial intelligence, artificial intelligence, machine learning, explainability.

I. INTRODUCTION

With the ever-growing adoption of Machine Learning (ML), the potential of fraudulence in highly complex ML models has recently raised significant concerns, leading to a surge of research in eXplainable Artificial Intelligence (XAI). Despite achieving promising results, XAI, a technology that focuses on elucidating opaque ML models, still faces significant challenges that need to be addressed [1], [2].

One major problem pertains to the quantitative measurement of the trustworthiness of an XAI explanation [3]. In the industry context, this is especially pertinent, as ML developers require the XAI not only to validate their models but also to demonstrate their reliability to potential users. Traditional performance criteria are insufficient in this regard; an additional focus on user trust is needed. While the XAI continues to maintain its prominence in the assessment of various ML models, it is noteworthy that a deficiency persists in the current form of quantitative evaluation frameworks, where the majority of research addressing this issue either struggles with the complexity of extensive human surveys or resorts to proxy methods based on the author's subjective assessment [4], [5], both of which are unsuitable in the company setting. This deficiency consequently renders all current evaluations reliant on qualitative measurements, thereby lacking practicability in the evaluation process.

For instance, in the renowned LIME paper [6], to validate the ability of LIME explanations to indicate better classifiers, the authors introduced a human experiment by strategically or randomly showing explanations from two classifiers to

The associate editor coordinating the review of this manuscript and approving it for publication was Xiaojie Su.

FIGURE 1. A demo of two different models' prediction reasoning.

human participants. The authors deliberately violate one classifier by training it with a biased dataset, while the other classifier is trained on the real dataset. The participants were then asked to select better explanations. Most people correctly select the explanations from the real model. Accordingly, they concluded the usefulness of their method for benefiting non-experts in choosing better models. In their qualitative evaluation approach, several issues emerge: first, the underlying factors that drive participants to select certain explanations are not clear. Second, the degree to which one explanation is superior to another remains undefined. Third, the transition from evaluating individual explanations to assessing the entire model seems unreliable.

To figure out answers of these issues and propose our own evaluation framework, we replicated the human experiments in [6] by presenting explanations from two distinct models, as illustrated in Fig. 1.

This time, the ML's task was to predict whether an individual has a high or low income based on their personal information. Notably, the majority of participants (37 out of 46) perceived model B as more reliable. We proceeded to inquire about the reasons behind their selections. The participants' responses revealed two primary criteria, which we summarize as follows:

Consistency with common sense in the direction of feature importance: Participants believed that certain features should logically have a positive or negative impact on predicting a high salary. For instance, they expected a bachelor's degree to positively influence the prediction of a high salary rather than a low salary.

Consistency with common sense in the relative importance of features: Participants expressed that the order of feature importance, based on their absolute values, should align with their intuitive understanding. For instance, they believed that gender should not be more influential in predicting one's salary than their level of education.

Based on the responses we have obtained, it seems feasible to gather human perspectives on ''common sense. We can then incorporate them into a framework that can be compared to the explanations derived from the XAI. Furthermore, the explainability of these explanations can be quantified based on their similarity to this ''common sense''. Also, the overall explainability of the model can be assessed by collectively evaluating each individual explanation.

Nevertheless, it's important to consider additional factors when evaluating explanations. According to Carvalho et al., there are mainly three goals for better explainability: (1) user's understandability, (2) user's interpreting efficiency, and (3) XAI approximating accuracy [7]. While understandability is strongly related to the similarity between human cognition and the XAI explanations as we mentioned previously, we also want to give solutions to quantify the user's interpreting efficiency and the XAI approximating accuracy.

In summary, our proposed quantification method is as follow: To assess the understandability of the explanation,

we propose a method to map human cognition to a linear model and then compare this human-defined model with the locally approximated model obtained from SHAP. For interpreting efficiency, we measure the number of cognitive chunks and feature diversity to estimate the efficiency of human understanding. Finally, the difference between the XAI local prediction and the model prediction is used to measure the XAI approximating accuracy. Although we aim to create a universal and model-agnostic method to quantify the explainability of all ML models, considering the variety of ML tasks, we will only focus on classification problems related to the tabular dataset in this research.

Together with the performance metrics of ML models, there are many potential applications of this explainability metric. For instance, in the problem of model selection, if several models perform very closely in terms of performance metrics, we can choose the model with the highest explainability for better explanations, which is extremely important in industries where data scientists need to advertise their models to end-users. Another application is a potential probe to monitor the explainability change of a model after updates caused by either a different training process or a change of the training set. To the best of our knowledge, no previous research has used human strategies as a reference to measure the explainability of ML models. This makes our study probably the first to address this problem.

Therefore, this paper's main contributions are as follows:

1) We introduce a novel framework that quantifies the explainability of ML models by SHAP and references to human strategy.
2) Our research uncovers various patterns within the ML training process. These observations emerged during the implementation of our method and provide insights into the dynamics of explainability during model training.
3) We critically examine and reevaluate several widespread assumptions in the XAI domain, particularly in practical applications. Our findings challenge the conventional understanding and present new perspectives on the explainability of ML models.

The remainder of this paper is as follows. In Section II, we present some of the existing work related to the field of explainability's quantification. In Section III, we formally define the objective of this work and explain in detail our proposed method. In Section IV, we show a real-world example obtained by running a human survey and discuss in detail the obtained results. In Section V we discussed several potential flaws in today's ML selection process. Finally, in Section VI, we conclude this paper.

II. RELATED WORK

In this section, we provide an overview of the local model-agnostic XAI and discuss some related research on quantifying the ML's explainability.

A. LOCAL MODEL-AGNOSTIC XAI

Local model-agnostic XAI refers to a category of XAI techniques. As the term "agnostic" suggests, these techniques do not rely on the internal structure or specifics of the underlying ML model but can be applied to any type of model that maps input to output. As the term "local" suggests, this type of XAI can only explain individual prediction rather than the general strategy used by the ML models.

One common approach used in local model-agnostic XAI is to generate explanations in the form of feature importance. These explanations are typically generated by permuting the input and observing how the prediction changes, which highlights the contribution of each input feature to the final prediction.

One popular local model-agnostic XAI method is SHAP, introduced by Lundberg and Lee in 2017 [8]. SHAP is based on Shapley values [9] from cooperative game theory and provides explanations for individual predictions by attributing the contribution of each feature to the final output. It offers different modules that can be applied to various types of ML models, such as kernel-SHAP for all types of models, deep-SHAP for deep learning models, and tree-SHAP [10] for tree-based models.

Other local model-agnostic XAI techniques, such as LIME [6], also operate on the same principle of providing local explanations by approximating the model's behavior around the specific instance being explained.

B. QUANTIFICATION OF EXPLAINABILITY

According to Arrieta et al., explainability can be regarded as an active characteristic of a model, denoting any action or procedure taken by a model to clarify or detail its internal function [11]. In contrast, interpretability refers to a passive characteristic of a model referring to the level at which a given model makes sense to a human observer. In our study, we consider explainability as an attribute of ML models, while interpretability describes the extent of human comprehension. Accordingly, a model with low explainability always lacks the capability to be interpreted and trusted by humans, and a model with high explainability can easily be interpreted and trusted by humans.

Since explainability is highly related to both ML structures and humans, the study of quantifying explainability can also be implemented from both directions. Some studies like [4] attempted to quantify the complexity of arbitrary ML models based on functional decomposition, which focuses on three criteria: the number of features used, the interaction strength, and the main effect complexity. Moreover, Mohseni et al. have addressed the inversely proportional relationship between a model's complexity and its explainability [12]. Others also suggested that some models like decision trees or linear models are intrinsically explainable models, while ensemble models and deep learning models are not, also known as black-box models [13]. Although the taxonomy seems well-founded, it still lacks practical value in the real-

world scenario, where different models' explainability cannot be directly compared due to different explanation methods. Meanwhile, models claimed to be more explainable due to their structures usually lack human subjective opinions, leaving these methods unsupported in terms of real-world usage [3].

On the other hand, starting from the human side, some studies focused on collecting human subjective reviews to measure the explainability of explanations. For example, Schmidt and Biessmann used the information transfer rate to quantify the explainability of any given explanation [5]. This transfer rate is measured by the human interaction time and the similarity between the human's prediction and the model's prediction. Nevertheless, Mohseni and Block [14] created multi-layer human attention masks aggregated from multiple human annotators and compared them with the model saliency explanations obtained by Gradient-weighted Class Activation Mapping (Grad-CAM) [15] and LIME [6]. Their idea is most similar to ours in the literature. As their study mainly focuses on images, they mention several limitations of the study, including the lack of reproducibility and the high cost of human annotation. These problems are, however, relatively easy to address for tabular datasets, which we will further elaborate in the next section.

To conclude, the review of previous research shows the dissonance between subjective and objective measurements of explainability. Although some attempts have been implemented to cover the gap, they seem rather unpractical in real-world scenarios.

III. PROPOSED APPROACH

First, we should define the problem setup. Let $\mathcal{A}$ denote an ML algorithm, and λ represent a parameter set. For any $\mathcal{A}$ with its parameters instantiated to λ, denoted as $\mathcal{A}_\lambda$, our target is to quantify the explainability of such model $\mathcal{A}_\lambda$. This explainability will be based on the model's ability to achieve the following goals: (1) The strategy of such a model extracted by an XAI method should make sense to the human. (2) The XAI approximation should be accurate. (3) The explanation itself should enable the human to make a quick judgment. Accordingly, we have created a human-grounded quantification method to quantify such explainability.

A. A MATHEMATICAL MODEL FOR HUMAN STRATEGY

Although it is nearly impossible to construct a mathematical model that can represent human cognition for a given task globally, we find it possible to represent it locally. For instance, given the task of predicting whether a person's salary is over 50,000 USD, it is hard to tell how education would globally influence the prediction. However, locally most people can annotate the possible impact if a person has a tertiary education level. Therefore, it is possible to use a linear combination with the weights and their associated features to locally represent a strategy.

Let h be the prediction model by humans. $h(x)$ is a local prediction based on a single input x. Any binary human prediction can be approximated by the following equation:

$$h(x) = \text{sign}(\sum_{i=1}^{M} \phi_i x_i). \tag{1}$$

Here, $x_i \in \{0, 1\}^M$, and M is the number of all features' values. $\phi_i \in [-1, 1]$, indicates the possible impact of one feature value on the final prediction. Therefore, extracting the human strategy translates into deciding ϕ_i for any possible feature's value in a given dataset.

It is also important to clarify the difference between the terms ''features'' and ''feature values'' here. The term ''feature'' indicates the piece of information itself regardless of the value it takes, whereas the term ''feature value'' indicates the instantiation of that information by giving it a value. For instance, *[education]* is a feature, while *[education = doctor]* is a possible value for this feature.

According to our human survey, individuals typically process these two terms at different levels. When evaluating values from the same feature, humans usually have a certain ranking in mind. However, comparing values belonging to different features proves to be more challenging. For example, participants show a strong preference for a tertiary education level over a high school education level in deciding the salary but struggle when asked to compare a tertiary degree with a specific working-hours.

On the other hand, judging the importance of a feature itself is usually done by comparing it with other features. For instance, while it may be difficult to directly compare a specific education level with working hours, humans find it easier to determine that the education-level holds more significance than the working-hours when it comes to salary decisions. Therefore, we believe that the annotation of the features and the feature values should be implemented separately. We will explain the annotation part in more detail in the next subsection. Moreover, since annotating all possible numerical values is impossible, we will use bins to discretize the numerical features.

B. ANNOTATION FOR FEATURE WEIGHTS

Based on the issues addressed in the previous section, we have developed a Graphical User Interface (GUI) to effectively gather people's opinions on different feature values.

To streamline the process, we have implemented a selection box that prompts users to focus on one feature at a time, while leaving the assessment of interrelationships between different features to global importance annotations. An illustrative example is depicted in Fig. 2 and Fig. 3.

Another challenge we encountered during our experiment was the difficulty for humans to annotate all feature values from scratch. For instance, annotators may have the idea that individuals with a master's degree generally earn more than high school graduates. However, determining the exact extent of this difference proves to be extremely challenging. To overcome this hurdle, instead of requiring users to annotate everything from scratch, we provide default values

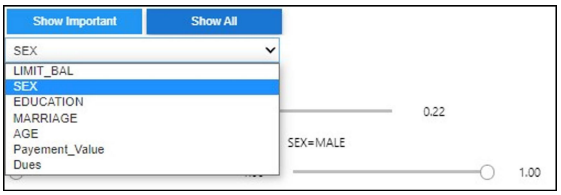

FIGURE 2. Demo on GUI about selecting features.

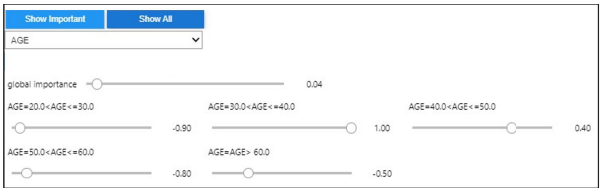

FIGURE 3. Demo on GUI about feature annotation.

by calculating the correlations between each feature value and the target class using Cramér's V [16]. Thus, users' task is not to establish these correlations but to modify them if they find them implausible or irrelevant.

Now, let $\theta_d \in [0, 1]$ denote the global importance of the d-th feature, and ρ_d^p denote the feature importance of d-th feature's p-th value. The ϕ_i of the human model is then obtained by the following equation:

$$\phi_i = \theta_d \times \rho_d^p. \tag{2}$$

The index i here indicates an arbitrary order of all feature values.

C. LOCALLY ABSTRACT ML'S STRATEGY

To make the human strategy and model strategy comparable, we also need to find a way to approximate any ML strategy to a linear model locally. In that sense, any additive feature attribution method [8] is applicable. Let k be the original prediction model and l be the explanation model. For any predication $k(x_i)$ based on a single input x_i, local explainable methods try to ensure $l(z_i) \approx k(z_i)$, when $z_i \approx x_i$. Therefore, the explanation by all additive feature attribution methods can be represented by the following equation:

$$l(z_i) = \xi_0 + \sum_{i=1}^{M} \xi_i \cdot z_i, \tag{3}$$

where $z_i \in \{0, 1\}^M$, and M is the total number of feature values. ξ_i is the weight of i-th feature value and ξ_0 is a possible bias term for the linear model.

In eq. (1) and eq. (3), ϕ_i and ξ_i both attribute an effect to a specific feature value. These effects, indicating how models and humans utilize a specific feature value to make a prediction, are therefore directly comparable. In this research we use SHAP [8] to showcase our algorithm.

We chose SHAP over other local model-agnostic explanation methods for two primary reasons. Firstly, SHAP's results are grounded in the calculation of the Shapley value, which is based on a robust mathematical foundation. This ensures that any variations observed when adjusting SHAP's parameters are due to the approximation method rather than the underlying mathematics. This aspect of SHAP offers a significant advantage over methods like LIME, which constructs a lasso model to approximate feature importance but lacks the same mathematical rigor [13]. Secondly, SHAP stands out as an integrated tool within the field of Explainable AI (XAI). It incorporates several cutting-edge projects to enhance the accuracy and efficiency of its explanations. For instance, SHAP includes specialized modules tailored for different model types, such as tree-SHAP for tree-based models deep-SHAP for deep learning models, and linear-SHAP for linear models. These features make SHAP a more comprehensive and reliable choice for our research purposes.

D. QUANTIFICATION OF EXPLAINABILITY OF A SINGLE PREDICTION

To measure the explainability of any given model $\mathcal{A}_\lambda$, we need to first quantify the explainability of any single prediction $k(x_i)$. As we stated previously, this explainability will be calculated based on the explanation's similarity, accuracy, diversity, and simplicity.

1) SIMILARITY BETWEEN A MODEL AND THE HUMAN STRATEGY

As we mentioned in Subsections III-A and III-C, for a given example $x_i \in X$, we will have multiple humans annotating feature importance values denoted as ϕ_i and multiple XAI feature importance values denoted as ξ_i. To better represent these values, we rearrange the index by introducing two new matrices, namely C_{ij} by changing the order of ϕ_i and W_{ij} by changing the order of ξ_i. Here, i is the index of the example, and j is the index of the features. The similarity between human cognition and model strategy will then be denoted as:

$$E_s(x_i) = S_c(C_{ij}, W_{ij}), \tag{4}$$

where $S_c(C_{ij}, W_{ij})$ indicates the cosine similarity of two vectors.

2) ACCURACY OF APPROXIMATION

In Subsection III-C, we mentioned that all local explainable methods try to ensure $l(z) \approx k(z)$ when $z \approx x$. Therefore, the approximation accuracy can be measured by the difference between $l(z)$ and $k(z)$. Here for a given example $x_i \in X$, the accuracy of the approximation can be denoted as the following equation:

$$E_a(x_i) = |l(x_i) - k(x_i)|. \tag{5}$$

3) FEATURE DIVERSITY (NON-REDUNDANCY)

How feature diversity will influence humans in explanation is when several features that the model considered necessary are very similar. For example, suppose we need to explain a model trained by a dataset lacking feature engineering,

containing over 50 features, many of which are very similar or highly correlated to a point where we can derive one from another. A good model strategy should focus on every aspect of an example from the users' perspective. Therefore, to measure the goodness of explanation, we should also create a mechanism to reward feature diversity.

To achieve that, we need first to let domain experts group features based on their similarities. These groups will then be denoted as F_g, where each F_g contains a certain number of features f_n^g. That is to say, different groups do not necessarily have the same number of features, and f_n^g simply refers to the number of features belonging to the g-th group. Suppose the number of features is N_f, that of groups is N_g, and $W_{f_n^g}$ indicates the feature importance of f_n^g, t is the user-defined threshold. For example x_i, we have the following equation to measure the feature diversity:

$$E_d(x_i) = \frac{\sum_{g=1}^{N_g} \mathbb{1}[\exists f_n^g \in F_g, |W_{f_n^g}| > t]}{N_g}. \tag{6}$$

The underlying idea behind this equation is as follows: for features that are deemed highly important by XAI (using a user-defined threshold t to determine high importance), we calculate the number of groups to which these features belong. If they belong to a diverse range of groups, we consider the feature diversity to be high. Conversely, if they all belong to a small number of groups, the diversity will be low.

4) SIMPLICITY

As many researchers have already addressed in the literature, the size of the explanation, namely the number of features used in our case, will influence the user's interpretability. Therefore, models incorporating fewer features yet keeping the same accuracy should also be rewarded in a model selection scenario. The simplicity of the explanation based on an example x_i will be quantified as follows:

$$E_e(x_i) = \frac{1}{N_f}. \tag{7}$$

To conclude this section, the explainability of a single prediction $f(x_i)$, can be quantified by combining all four criteria:

$$E(x_i) = \omega_1 E_s(x_i) + \omega_2 E_a(x_i) + \omega_3 E_d(x_i) + \omega_4 E_e(x_i), \tag{8}$$

where ω_1, ω_2, ω_3, and ω_4 are weights given to each component of the explainability, highlighting its importance for target users to understand an explanation.

5) MULTI-CLASS CASE

Previously, we only introduced how to quantify the explainability of binary classification problems. However, since we want to propose a method covering all classification problems related to the tabular dataset, a technique for the multi-class problem is needed.

In SHAP [8], all N-class explanations can be separated into N-binary explanations, where $\sum_{n=1}^{N} l^n(z_i) = 1$. $l^n(z)$ represents the model prediction of the possibility of class n Accordingly, N sets of feature importance will be provided. Therefore, for an N-class classification problem, the explainability of a prediction based on x_i can be quantified as follows:

$$E(x_i) = \frac{\sum_{n=1}^{N} E^n(x_i)}{N}. \tag{9}$$

E. QUANTIFICATION OF MODEL'S EXPLAINABILITY

Accordingly, for any given model $\mathcal{A}_\lambda$, and a specific test dataset D_{test} for testing the explainability, whose size (cardinality) is S_t. The explainability of $\mathcal{A}_\lambda$ should not only be measured by the average value of all prediction's explainability but also by its standard deviation, which can be calculated using the following equation:

$$E(\mathcal{A}_\lambda, D_{\text{test}}) = \frac{\sum_{i=1}^{S_t} E(x_i)}{S_t} - \omega_5 \cdot \text{std}(E(D_{\text{test}})), \tag{10}$$

where $\text{std}(E(D_{\text{test}}))$ is the standard deviation of the vector $E(D_{\text{test}})$, which includes all single explainabilities of the examples in the set D_{test}.

IV. SIMULATED EXPERIMENT

In this section, we conducted simulated user experiments to assess the effectiveness of our explainability metric in capturing real human ideas. To achieve this, we trained multiple models to make predictions, which were then explained using the SHAP technique and evaluated using our explainability metric.

It is worth noting that even when presented with the same example, different models may generate distinct prediction strategies. Thus, the objective of our human survey was to expose participants to these diverse explanations and examine whether their preferences, in terms of trust, aligned with our explainability metric. If the human participants consistently favored explanations with higher explainability, it would serve as evidence supporting the utility of our method.

The simulated experiment consists of three primary stages: 1) Human strategy modeling, 2) Example selection, and 3) Human survey. When given a specific task, we start by constructing a human model using annotations from domain experts or volunteers. Next, we utilize a specific algorithm to choose representative predictions from various ML models and explain them using the SHAP technique. Finally, we present different pairs of explanations to humans and ask for their preferences, which helps us evaluate the feasibility of our metric.

In the subsequent subsections, we will first introduce the dataset we used in this survey and then present a detailed description of each stage, elucidating the methodology and procedures employed in our study.

A. DATASET

In terms of dataset selection, considering the majority of participants in our survey are students, our focus is on datasets that are widely recognized and understood by the

general public. This strategy is intended to reduce potential misunderstandings that could arise from the use of highly specialized or technical data. Therefore, we carefully chose three distinct datasets for our experimental analysis. These datasets include the census income dataset [17], the default prediction dataset [18], and the laptop price dataset.[1]

In the following subsections, we provide a comprehensive description of each dataset to offer a clear understanding of their characteristics and relevance to our study.

1) INCOME DATASET

The primary objective of the census income dataset is to predict whether an individual's income exceeds $ 50,000 USD. The dataset comprises over 32,000 examples, each containing 14 distinct features. For the purpose of our human survey, we have excluded certain sensitive attributes such as race and nationality to ensure fairness and prevent bias. Additionally, we performed feature engineering to consolidate related features. For instance, we combined capital gain and loss to derive the pure capital gain. Consequently, the final training dataset consisted of 8 features, namely age, family members, working hours per day, occupation, gender, marital status, education level, and pure capital gain.

2) THE DEFAULT PREDICTION

The primary objective of this dataset is to predict whether an individual will experience a credit card default based on their credit history spanning five months. The dataset encompasses various features, including personal information and the individual's default history over the past five months. These features provide valuable insights into the individual's financial behavior and patterns, aiding in the prediction of credit card defaults.

3) THE LAPTOP PRICE

This dataset comprises a comprehensive inventory of various computer components, including their respective brands. The objective of this dataset is to predict whether a given computer can be classified as high-end or not. By examining the specific components and brands associated with each computer, the model can make an informed prediction regarding its classification.

In the subsequent subsections, we provide a detailed explanation of how we extract human ideas from specific tasks and utilize them to build the human model.

B. HUMAN STRATEGY MODELING

At this stage, our objective is to extract the human strategy employed in specific tasks and convert it into a linear model for comparison with the ML strategy. The overall process of this stage is depicted in Fig. 4.

We developed a Graphical User Interface (GUI) specifically designed for extracting strategies from human participants. This application serves four primary purposes:

[1]https://www.kaggle.com/datasets/muhammetvarl/laptop-price

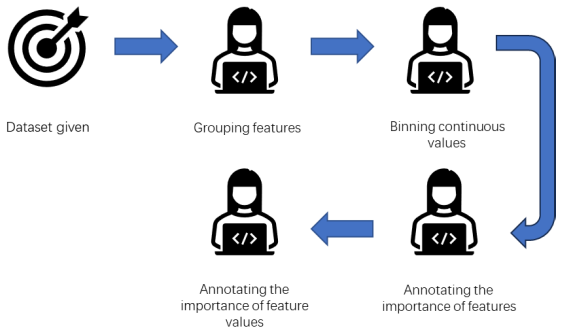

FIGURE 4. The pipeline of the human strategy modeling.

1) GROUPING FEATURES

When presented with a particular task and its associated dataset, we engage domain experts to categorize the different features into distinct groups. These categories play a crucial role in the subsequent computation of feature diversity. However, through effective feature engineering techniques, it is possible to render all the included features uncorrelated. Consequently, each category may encompass only one feature.

2) BINNING CONTINUOUS VALUES

Domain experts utilize the GUI to categorize all continuous values by examining the density function of each continuous feature. This step ensures that the generated bins do not contain an excessive or insufficient number of examples, thereby enhancing the accuracy and reliability of the subsequent analysis.

3) ANNOTATING THE IMPORTANCE OF FEATURES

Domain experts are tasked with annotating the relative importance of all features. For instance, in the adult income dataset, they would annotate features such as education, gender, occupation, age, and so on, based on their perceived significance.

4) ANNOTATING THE IMPORTANCE OF FEATURE VALUES

Within each feature, domain experts provide annotations for all the specific feature values. For example, under the "education" feature, they would assign importance values to feature values such as primary school, middle school, bachelor's degree, master's degree, and so on. This step allows for a more comprehensive understanding of the varying degrees of importance within a given feature.

As we mentioned in III-B, we have incorporated default annotation values based on the calculation of Cramér's V correlation [16] between each feature value and the target class, as depicted in Fig. 3. These default values serve as a starting point, providing annotators with a reference that incorporates information derived from the dataset itself. Annotators have the flexibility to modify these default values

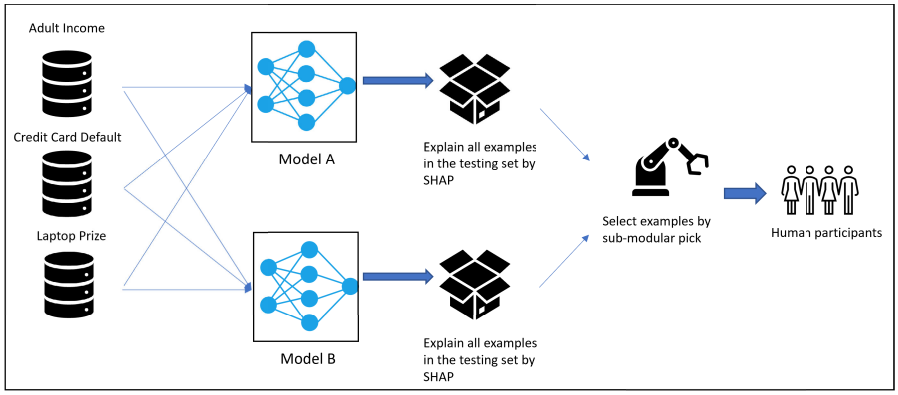

FIGURE 5. Pipeline of generating examples for human survey.

if they believe they are inappropriate or require adjustment based on their domain expertise.

C. EXAMPLE SELECTION

The example selection process is depicted in Fig. 5. As it is infeasible for users to review every single example in the dataset, we employ a strategy to select a representative subset of examples for human evaluation. It is important to note that the objective of our human survey is not to assess the quality of explanations but rather to measure the degree to which our method accurately captures real human ideas. In other words, if an explanation is deemed "bad" by our method, it should also be perceived as unsatisfactory by real humans. We summarize the selection strategy as sub-modular picking, referring to the selection strategy outlined in [6].

Unlike the sub-modular pick algorithm in [6], our algorithm addresses the problem of picking up a subset from the testing set that is suitable for comparison. Our goal of this survey is to see how the explainability values represent the real human idea but not to evaluate models in terms of explainability. We argue that a suitable subset for making comparisons should maintain the following criteria: (1) The subset selected by the algorithm should be within a human-defined budget B, which is the largest number of instances that human judges are willing to check. (2) The features included in the subset should be as diverse as possible - there is no use in asking participants to check similar examples over and over again. (3) The predictions' explanations of two models based on this subset should have a certain degree of difference - there are no benefits in asking participants to compare two explanations that look the same.

Given a set of instances X ($|X| = n$), we construct an $n \times d$ matrix O by one-hot encoding so that each column represents the absence or presence of one feature value. For each column j in O, let I_j denote the total appearance of one feature value. Further, for the two models that we use to select the subset, let E be an $n \times d$ explanation matrix of one model, E' be the explanation matrix of another, and D represents a difference vector between E_1 and E_2 with a size of d (each element in vector represents the cosine similarity between each row in E and E'). Finally, since we want to pick up a subset that can cover as many features' values as possible while maintaining a certain degree of difference, we formalize this coverage intuition in eq. (10):

$$c\left(V, O, I\right) = \sum_{j=1}^{d} \mathbb{I}\left[\exists j \in V : D_j > \sigma\right], \tag{11}$$

where c is a set function that, given O and I, if D_j is bigger than a threshold σ, computes the total appearance of features in a set V. The pick problem, defined in eq. (11), is to find the set V, where the new index i in V can achieve the highest marginal gain.

$$Pick\left(O, I\right) = \underset{V, |V| \leq B}{\operatorname{argmax}} \left(c\left(V \cup i, O, I\right) - c\left(V, O, I\right)\right). \tag{12}$$

We further define the procedure to greedily find the subset V in Algorithm 1.

Algorithm 1 Sub-Modular Pick

Require: Instance X, Explanations E, E', Budget B
  $O \leftarrow \text{onehot}(X)$
  for $j \in \left\{1 \ldots d'\right\}$ **do**
    $I_j \leftarrow \sum_{i=1}^{n} O_{ij}$
    $D_j \leftarrow S_c(E_j, E'_j)$
  end for
  $V \leftarrow \{\}$
  while $|V| < B$ **do**
    $V \leftarrow V \cup Pick(O, I)$
  end while

D. HUMAN SURVEY

For each dataset, we trained two different Dense Neural Networks (DNNs) with distinct structures to assess their performance in terms of explainability on the same examples. The first DNN was trained with a relatively simple structure, consisting of only one hidden layer with sixty nodes. In contrast, the second DNN was trained with six hidden layers, each containing sixty nodes.

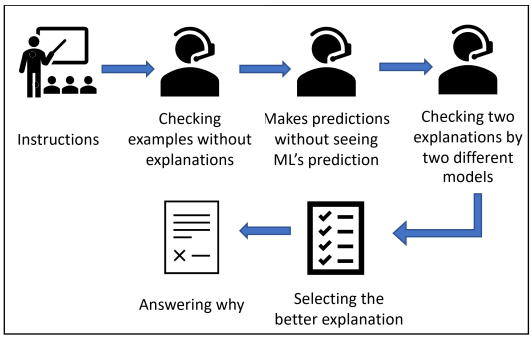

FIGURE 6. Pipeline of the 1st phase in human survey.

FIGURE 7. Showing one example in the dataset.

Both models utilized the Rectified Linear Unit (ReLU) activation function in the hidden layers, and the Sigmoid function as the activation function in the output layer. We evaluated the explainability metric using the same testing set, which accounted for 20% of the dataset. The performance metrics of the models are provided in Table 1.

Regarding the parameters for the explainability calculation, since the number of input features remains constant and all features belong to different groups, we set ω_3 and ω_4 to zero. Additionally, we assigned values of 1 to ω_1 and ω_2, and 0.6 to ω_5. These choices resulted in a range of explainability for the models approximately ranging from -1.6 to 1.0, while the range for the examples' explainability was approximately -1.0 to 1.0.

It is important to note that we use the term "roughly" because SHAP generates highly accurate local predictions, minimizing the effect of ω_2 near zero.

To assess whether a higher explainability value can indicate better interpretability for humans, we strategically selected five examples from each dataset. These examples were then predicted by all the trained models, and their predictions were explained using SHAP. Each pair of explanations was presented to the human participants, who were asked to compare them and evaluate their trustworthiness.

Finally, we recruited 46 participants with diverse backgrounds to participate in the human survey.

The human survey consists of two phases. In the first phase, participants are tasked with selecting their preferred model among the options provided. In the second phase, they are asked to assign a score ranging from 1 to 10 to specific explanations, indicating their perceived reasonability of the explanation's strategy.

The human survey follows the pipeline illustrated in Fig. 6. In the first phase, participants are initially presented with example information without any accompanying explanations, as shown in Fig. 7. This step allows participants to form their own judgments and construct their understanding without external influence. Subsequently, two predictions made by ML models are presented along with their corresponding explanations. Participants can compare the model's explanation with their own cognition and choose the preferred model. Finally, all participants are requested to provide reasons for their preference regarding a particular explanation.

In the second phase, participants are presented with ten predictions made by a single model, and each prediction is accompanied by an explanation. These explanations vary in terms of their level of explainability, ranging from high to low. Participants are then asked to assign scores to these ten explanations, indicating their perceived reasonability. Higher scores are assigned to explanations that make more sense to the participants.

TABLE 1. Model information and performance.

Model (Salary)	Metric		
	Explainability	Accuracy	F1
DNN-1	0.652	0.851	0.652
DNN-6	0.383	0.858	0.655
Model (Default)	**Metric**		
	Explainability	Accuracy	F1
DNN-1	0.232	0.702	0.700
DNN-6	-0.102	0.725	0.710
Model (Laptop)	**Metric**		
	Explainability	Accuracy	F1
DNN-1	0.417	0.873	0.885
DNN-6	0.270	0.880	0.885

E. SURVEY RESULTS

The results of the human survey are shown in Table 2, where the explainability of each prediction and the human choice ratio are shown.

The results show that all of the examples we picked have larger human preference ratios towards the explanations with higher explainability. Furthermore, the ratio difference of choice and value difference of explainability seems to be positively correlated - their χ^2 correlation values in the datasets in hand are 0.905, 0.956, and 0.711, respectively. However, the ratio of "hard to tell" in every example is much higher than we expected. We asked the participants why they consider it hard to tell which one is better. For all 153 answers

TABLE 2. Survey results of the comparison.

Salary	Metric		Human choice ratio		
	Exp(1-layer DNN)	Exp(6-layer DNN)	1 layer DNN	6-layer DNN	Hard to tell
1	0.92	0.47	63%	24%	13%
2	0.88	0.35	74%	15%	11%
3	0.75	0.42	61%	30%	9%
4	0.59	-0.05	74%	15%	11%
5	0.22	0.61	22%	67%	11%
Default	**Metric**		**Human choice ratio**		
	Exp(1 layer DNN)	Exp(6-layer DNN)	1 layer DNN	6-layer DNN	Hard to tell
1	0.91	-0.63	74%	22%	4%
2	0.63	-0.65	63%	20%	17%
3	0.63	-0.52	65%	24%	11%
4	0.53	-0.05	35%	35%	30%
5	0.27	-0.07	43%	39%	17%
Laptop	**Metric**		**Human choice ratio**		
	Exp(1 layer DNN)	Exp(6-layer DNN)	1 layer DNN	6-layer DNN	Hard to tell
1	0.90	0.31	61%	22%	17%
2	0.84	0.32	65%	20%	15%
3	0.75	0.21	57%	26%	17%
4	0.62	0.03	70%	22%	8%
5	0.51	0.02	43%	39%	17%

with ''hard to tell'', nearly 80% were given the reason ''Both are untrustworthy.''

In situations where there are significant differences in the explanations, most individuals can easily identify the better ones. However, as the level of explainability becomes similar, human decision-making becomes more diverse. It is therefore reasonable to assess how humans interpret the explanations provided by the two models, as the model with higher explainability is capable of generating more explainable explanations for most predictions derived from the same training set. Nonetheless, while people can generally recognize which explanations are reasonable and which are not, it can be challenging for them to determine the extent to which one explanation is better than another.

The results of the second phase also reflect this phenomenon, as depicted in Fig. 8, where participants generally agree that certain explanations are reasonable while others are not. However, for explanations with decent levels of explainability, it becomes difficult for humans to distinguish which one is better.

Another noteworthy observation from the second round's results is that as explainability decreases to a point near -1 (scaled to 0 in Fig. 8), human intuition regarding the reasonability of these explanations unexpectedly increases. Upon interviewing some participants, it was discovered that without any background information, they had no knowledge of the income levels associated with certain occupations or how specific features would influence the predictions, as they were students without expertise in the field. For example, in the 8th example shown in Fig. 8, the explanation indicates a significant positive impact of the feature value [occupation = protective service] (translated as policeman in our explanation), which is incorrect since we annotated this specific occupation with a negative impact. However, many participants believed that being a policeman is an indicator of high income, unaware of the fact that the average income of a service job in the USA was only $25k in 1994.

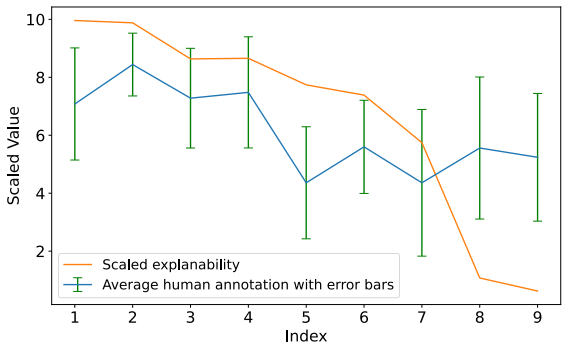

FIGURE 8. The average human annotation against the scaled explainability. Here, the explainability value of each explanation was scaled to match the range of human annotation, which is within 0 to 10.

V. DISCUSSIONS

A. WHY IS IT HARD TO COMPARE EXPLAINABILTY OF DIFFERENT MODELS

A commonly held belief in the literature is that simple models are inherently more explainable than complex models. Models such as decision trees, k-nearest neighbors, and linear models are often considered intrinsically explainable due to their human-interpretable rules.

However, in practical scenarios where vast datasets are used to train simple models, the strategies derived from such models may not always be easily interpretable. This is because simple models are limited in their ability to represent complex global strategies. For example, a simple rule list may not be sufficient to capture the intricate relationships within a dataset comprising 30,000 examples. Even if one attempts to construct such a rule list, it would likely be exceedingly

TABLE 3. A comparison between the random forest classifier and the decision tree one in terms of explainability, accuracy, and F1-score.

Model	Metric		
	Explainability	Accuracy	F1
Random Forest (RF)	0.436	0.799	0.858
Decision Tree (DT)	0.244	0.774	0.840

long, making it incomprehensible to humans despite being formulated in a human-understandable manner.

As a result, in practice, even for inherently explainable models, practitioners often utilize XAI technologies to simplify the rules and make them understandable to customers. Under these circumstances, we are curious about the differences in explainability between an explainable model and a black-box model.

To address this question, we trained two different models, a Decision Tree (DT) and a Random Forest (RF), using the same income dataset. Employing the default parameters provided by Scikit-Learn, the results of the two models are depicted in Fig. 9 and summarized in Table 3.

To our surprise, the performance of the DT and RF models does not differ significantly. However, the DT now exhibits a depth of 58 and maintains over 5000 leaves, which makes it practically impossible to transform into a reasonable set of rules that can be presented to customers. The resulting rule set would consist of thousands of confusing rules, some of which may even appear absurd.

In contrast, we utilized the SHAP method and our proposed approach to evaluate the two models. Interestingly, the RF model outperformed the DT in terms of explainability. This difference is clearly demonstrated in the Probability Density Function (PDF) depicted in Fig. 9, where we drew the PDF of all examples' explainability. We can tell the majority of predictions made by the RF model are more explainable compared to those made by the DT model - the green pattern has a larger density in the high-explainability region.

However, it is important to note that this experiment does not necessarily prove that the DT model is less interpretable than the RF model. This is because the explanations generated by SHAP are approximations and may not be as faithful as the direct rule lists generated by the DT model. Moreover, similar to the DT model, the RF model constructs multiple decision trees and employs a subset of them randomly (hence the term "Random" Forest) to generate predictions. These generated trees are even less comprehensive and understandable than the single DT's tree. Consequently, attempting to transform the RF model into a set of rules for customer presentation is even more futile, as these rules lack logical coherence and understandability.

Nevertheless, in practice, when faced with a complex DT or an intricate RF model, practitioners often resort to XAI methods such as SHAP. In this regard, we argue that practitioners should carefully consider which algorithms to employ, as so-called explainable models may not necessarily retain their explainability under such circumstances.

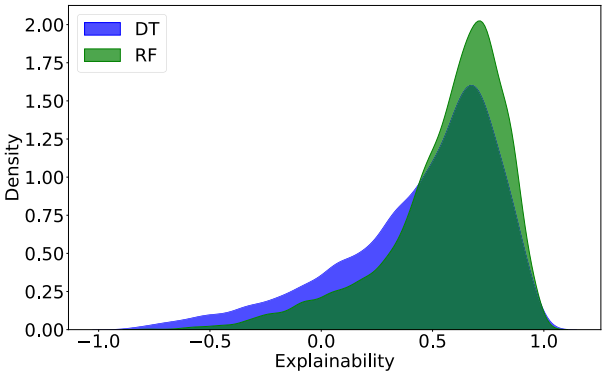

FIGURE 9. The PDFs of the two models' explainability: The Random Forest (RF) and the Decision Tree (DT).

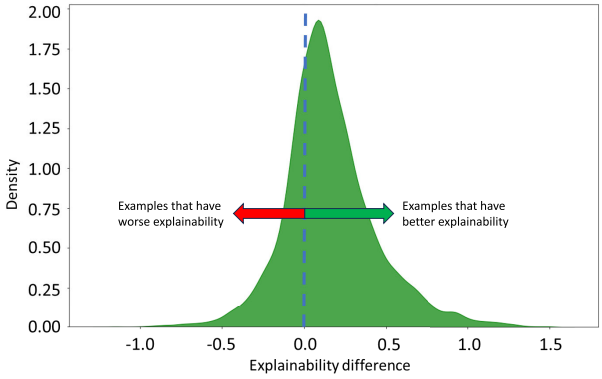

FIGURE 10. The PDF of the subtractions between the RF' explainability values and the DT's explainability values on the same example.

B. DIFFERENCE BETWEEN EXPLAINABILITY OF THE SINGLE PREDICTION AND THE MODEL

Most quantification methods using human references and local explanations try to strategically pick up a limited set of examples for human evaluation. The number of examples picked is restricted by how many examples humans are willing to check. For the most part, humans tend to prefer not to take a survey asking them to check thousands of examples. However, when evaluated by the human model, this problem does not exist.

In the previous experiment, depicted in Fig. 9, we suggest that the RF outperforms the DT in terms of explainability. This conclusion is drawn from the observation that the explainability values of the RF are more concentrated towards higher values. However, this does not imply that every prediction made by the RF is inherently more reasonable than those made by the DT.

To further demonstrate this, we calculate the differences in explainability between each prediction made by the RF and the corresponding prediction by the DT for the same example. The density plot of these explainability differences is presented in Fig. 10, where the blue dotted line separates the examples into two groups: better examples and worse

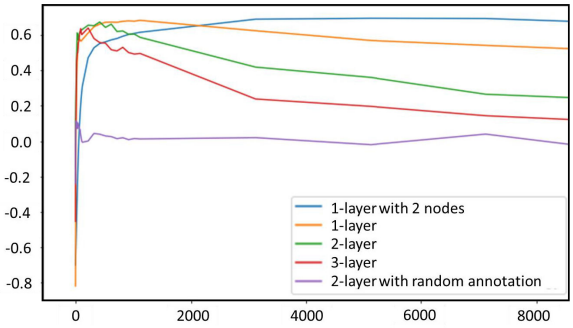

FIGURE 11. Explainability of neural networks with different structures per epoch.

examples. Intuitively, examples on the right side of the blue line indicate those with higher explainability in the RF, and vice versa.

While most predictions made by the RF exhibit higher explainability, it is important to note that around 20% of predictions made by the RF actually have lower explainability. Therefore, we argue that conventional methods of comparing models by selecting only a few examples are not reliable due to the diversity of explainability distribution. In other words, there is a possibility that the selected examples perform better with one model, explainability-wise, but when a sufficient number of examples are examined, another model may actually perform better.

C. EXPLAINABILITY IN THE TRAINING PROCESS

To observe how explainability evolves as the Neural Network updates its parameters, we employed four Neural Networks with distinct structures and monitored their explainability after each epoch, with the batch size set to the size of the training set.

The results, illustrated in Fig. 11, demonstrate an initial rapid increase in explainability for nearly all the structures. However, this upward trend eventually diminishes, and the explainability values start to decline beyond certain points.

Furthermore, while all the models (except the model with random annotation) are able to achieve a similar peak level of explainability, it is notable that models with more complex structures exhibit greater sensitivity to explainability. These models show faster improvement in explainability but also experience faster declines.

On one hand, the substantial increase in explainability during the early epochs can be attributed to the model initially establishing a set of general rules that align, to some extent, with human intuition regarding what factors are most important.

On the other hand, as the model continues to learn and adapt to the training data, it begins to develop more complex and customized rules that specifically fit the individual training samples or batches. Consequently, the model starts to diverge from the human annotations. In other words, the model learns very specific rules that are true for only a few samples in the current dataset, but these rules do not apply to the majority of samples, resulting in counter-intuitive explanations.

This phenomenon is further highlighted by the manner in which each model's explainability declines. Simpler networks have limited capacity to construct complex rules, causing them to adhere to rules that align with the majority of samples in the training set. Conversely, more complex networks have the ability to learn highly specific rules, leading them to overfit in terms of explainability even before they overfit in terms of loss.

To examine the influence of human annotations on the model's explainability, we monitored the explainability change of a 2-layer Deep Neural Network (DNN) with random annotations. These annotations were randomly generated by the computer. Under such circumstances, the explainability barely changes and remains close to zero, indicating a lack of meaningful correlation. This highlights the importance of having reasonable human annotations to effectively quantify the model's explainability. We speculate that if these random annotations were to become less random and more structured, we might observe the corresponding curves approaching those of other models.

Our experiment on explainability change demonstrates the impact of the training process on explainability, an aspect that is often overlooked by many ML practitioners. Therefore, it is crucial to monitor these changes in explainability in addition to assessing model accuracy and loss.

VI. CONCLUSION

In this paper, we have presented a novel solution to quantify the explainability of any ML classifier by leveraging human strategy. Our approach involves domain-knowledgeable users annotating all possible feature values in a given dataset, thereby constructing a local linear model that approximates human strategy. This linear model is then compared to the linear model generated by SHAP based on various ML models. We have also incorporated several well-studied criteria of explainability from previous works, allowing users to choose the criteria that align with their preferences.

To validate the effectiveness of our explainability metric, we conducted a human survey involving 46 participants. The results demonstrated that our metric approximated the human interpretability of explanations effectively. During the implementation of our method, we observed several notable patterns in the ML training process. One key observation was that explainability often follows a distinct trajectory during training. Initially, it increases rapidly but starts to decline after reaching a certain threshold. This insight underscores the need to consider both the architecture and the training dynamics of ML models when evaluating them based on explainability.

Our method also allowed us to challenge and reassess some prevalent beliefs in practical settings. For instance, our experiments revealed that models traditionally labeled as 'explainable' may not always surpass 'unexplainable'

models in terms of actual explainability. In one experiment, using SHAP to interpret both a decision tree and a random forest trained on the same dataset, we found scenarios where the random forest was more explainable. Furthermore, we investigated the reliability of using a small number of examples to gauge a model's overall explainability. Our findings indicate a wide variability in the explainability of individual predictions. Interestingly, models with lower overall explainability sometimes produced predictions with higher explainability than those from more explainable models. This variation suggests that deciding a model's explainability based on a limited set of examples can be misleading.

However, it is important to acknowledge several limitations that should be taken into consideration when interpreting the results. These limitations include the followings:

1) In our human survey, we did not incorporate all the criteria outlined in our mathematical model, though many of these have been thoroughly investigated in previous studies.
2) Our proposed method is tailored specifically for tabular datasets and classification tasks, limiting its applicability to other types of data, such as text or images.
3) The effectiveness of our quantification process is highly dependent on the quality of human annotations. We demonstrated in one of our experiments how random annotations can be ineffective in accurately measuring explainability. Therefore, developing an annotation strategy that gains wide acceptance among potential users is critical.

Despite these limitations, we believe that our method is robust and applicable to a wide range of classification tasks in the field of data science. It is crucial to highlight the potential risks associated with solely relying on accuracy during the ML training process. By neglecting the importance of explainability, ML models may struggle to maintain the trust of end-users. Our work serves as a valuable contribution to addressing this issue and emphasizes the need to consider explainability alongside accuracy in the development and deployment of ML models.

REFERENCES

[1] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, *arXiv:1702.08608*.
[2] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue*, vol. 16, no. 3, pp. 31–57, Jun. 2018.
[3] S. Verma, A. Lahiri, J. P. Dickerson, and S.-I. Lee, "Pitfalls of explainable ML: An industry perspective," 2021, *arXiv:2106.07758*.
[4] C. Molnar, G. Casalicchio, and B. Bischl, "Quantifying model complexity via functional decomposition for better post-hoc interpretability," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Cham, Switzerland: Springer, 2019, pp. 193–204.
[5] P. Schmidt and F. Biessmann, "Quantifying interpretability and trust in machine learning systems," 2019, *arXiv:1901.08558*.
[6] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' explaining the predictions of any classifier," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.
[7] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, Jul. 2019.
[8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, p. 4765.
[9] L. Shapley, "17. A value for n-person games," in *Contributions to Theory Games (AM-28)*, vol. 2. Princeton, NJ, USA: Princeton Univ. Press, 2016, pp. 307–318.
[10] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Mach. Intell.*, vol. 2, no. 1, pp. 56–67, Jan. 2020.
[11] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.
[12] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable AI systems," *ACM Trans. Interact. Intell. Syst.*, vol. 11, nos. 3–4, pp. 1–45, Dec. 2021.
[13] C. Molnar. (2018). *A Guide for Making Black Box Models Explainable*. [Online]. Available: https://christophm.github.io/interpretable-ml-book
[14] S. Mohseni, J. E. Block, and E. Ragan, "Quantitative evaluation of machine learning explanations: A human-grounded benchmark," in *Proc. 26th Int. Conf. Intell. User Interface*, Apr. 2021, pp. 22–31.
[15] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[16] H. Cramér, *Mathematical Methods of Statistics*, vol. 26. Princeton, NJ, USA: Princeton Univ. Press, 1999.
[17] K. Ron, "Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid," in *Proc. Kdd*, vol. 96, 1996, pp. 202–207.
[18] I.-C. Yeh and C.-H. Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 2473–2480, Mar. 2009.

ZHAOPENG LI received the B.E. degree in communication engineering from the University of Science and Technology, Beijing, China, in 2019, and the M.E. degree from Keio University, Japan, in 2023, where he is currently pursuing the Ph.D. degree. His research interests include explainable artificial intelligence and data science.

MONDHER BOUAZIZI (Member, IEEE) received the B.E. degree in communications from SUP-COM, Carthage University, Tunisia, in 2010, and the M.E. and Ph.D. degrees from Keio University, in 2017 and 2019, respectively. He was a Telecommunication Engineer (access network quality and optimization) with Ooredoo Tunisia (Ex. Tunisiana), for three years. He is currently a Specially Appointed Assistant Professor with the Ohtsuki Laboratory. He has published several journal and international conference papers. He has engaged in research on machine learning, deep learning, data mining, sensors, and signal processing. He is a member of the Association for Computing Machinery (ACM) and the Institute of Electronics, Information and Communication Engineers (IEICE). He received the Telecommunications Advancement Foundation Student Award, in 2016; the IEEE/ACM ICSIM, in 2021; the IEEE APPC 2021 Best Paper Award; and the A3 Workshop 2021 Best Presentation Award.

TOMOAKI OHTSUKI (Senior Member, IEEE) received the B.E., M.E., and Ph.D. degrees in electrical engineering from Keio University, Yokohama, Japan, in 1990, 1992, and 1994, respectively. From 1994 to 1995, he was a Postdoctoral Fellow and a Visiting Researcher of electrical engineering with Keio University. From 1993 to 1995, he was a Special Researcher of Fellowships of the Japan Society for the Promotion of Science for Japanese Junior Scientists. From 1995 to 2005, he was with the Science University of Tokyo, Tokyo, Japan. In 2005, he joined Keio University, where he is currently a Professor. From 1998 to 1999, he was with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA, USA. He has engaged in research on wireless communications, optical communication, signal processing, and information theory. He has published more than 205 journal articles and 415 international conference papers. He is a fellow of IEICE and a member of the Engineering Academy of Japan. He was a recipient of the 1997 Inoue Research Award for Young Scientist; the 1997 Hiroshi Ando Memorial Young Engineering Award; the Ericsson Young Scientist Award, in 2000; the 2002 Funai Information and Science Award for Young Scientist; the IEEE the first Asia–Pacific Young Researcher Award, in 2001; the 5th International Communication Foundation Research Award; the 2011 IEEE SPCE Outstanding Service Award; the 27th TELECOM System Technology Award; the ETRI Journal's 2012 Best Reviewer Award; and the Best Paper Award at 9th International Conference on Communications and Networking, China, in 2014. He served as the Chair for IEEE Communications Society and the Signal Processing for Communications and Electronics Technical Committee. He has served as the General Co-Chair, the Symposium Co-Chair, and the TPC Co-Chair for many conferences, including IEEE GLOBECOM 2008, SPC, IEEE ICC 2011, CTS, IEEE GLOBECOM 2012, SPC, IEEE ICC 2020, SPC, IEEE APWCS, IEEE SPAWC, and IEEE VTC. He gave tutorials and keynote speeches at many international conferences, including IEEE VTC, IEEE PIMRC, and IEEE WCNC. He was the Vice President and the President of the Communications Society of IEICE. He served as a Technical Editor for *IEEE Wireless Communications Magazine* and an Editor for *Physical Communications* (Elsevier). He is serving as an Area Editor for IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY and an Editor for the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS. He is a Distinguished Lecturer of IEEE.

MASAKUNI ISHII received the B.S. degree in mathematics and computer science from The University of Arizona, AZ, USA, in 2007, and the M.E. degree in computer engineering from Keio University, Yokohama, Japan. He is a Senior Researcher with Nippon Telegraph and Telecommunication Corporation (NTT), Tokyo, Japan. His research interests include data sciences, security, and psychology.

ERI NAKAHARA received the B.S. degree in contemporary society from Kyoto Women's University, Japan, in 2017, and the M.S. degree from the Graduate School of Information Science, Nara Institute of Science and Technology, Japan. She joined Nippon Telegraph and Telephone Corporation Laboratories, in 2019. She has been a member of the Computer and Data Science Laboratory, since 2021.