Journal of Artificial Intelligence Research 75 (2022) 833-855 Submitted 06/2022; published 11/2022

Interpretable Local Concept-based Explanation with Human Feedback to Predict All-cause Mortality

Radwa Elshawi RADWA.ELSHAWI@UT.EE
Institute of Computer Science
Tartu University, Estonia

Mouaz H Al-Mallah MAL-MALLAH@HOUSTONMETHODIST.ORG
Houston Methodist DeBakey Heart & Vascular Center
Houston, TX, USA

Abstract

Machine learning models are incorporated in different fields and disciplines in which some of them require a high level of accountability and transparency, for example, the healthcare sector. With the General Data Protection Regulation (GDPR), the importance for plausibility and verifiability of the predictions made by machine learning models has become essential. A widely used category of explanation techniques attempts to explain models' predictions by quantifying the importance score of each input feature. However, summarizing such scores to provide human-interpretable explanations is challenging. Another category of explanation techniques focuses on learning a domain representation in terms of high-level human-understandable concepts and then utilizing them to explain predictions. These explanations are hampered by how concepts are constructed, which is not intrinsically interpretable. To this end, we propose Concept-based Local Explanations with Feedback (CLEF), a novel local model agnostic explanation framework for learning a set of high-level transparent concept definitions in high-dimensional tabular data that uses clinician-labeled concepts rather than raw features. CLEF maps the raw input features to high-level intuitive concepts and then decompose the evidence of prediction of the instance being explained into concepts. In addition, the proposed framework generates counterfactual explanations, suggesting the minimum changes in the instance's concept-based explanation that will lead to a different prediction. We demonstrate with simulated user feedback on predicting the risk of mortality. Such direct feedback is more effective than other techniques, that rely on hand-labelled or automatically extracted concepts, in learning concepts that align with ground truth concept definitions.

1. Introduction

Machine learning (ML) models have proven to be successful in many application domains, including financial systems, advertising, marketing, criminal justice, especially with the advent of deep learning (Sakr & Zomaya, 2019; Saggi & Jain, 2018; Wang, Pan, He, Huang, Wang, & Tu, 2020). The study of personalized agents, recommendation systems, and critical decision-making tasks (e.g., medical analysis) has added to the importance of machine learning interpretability and artificial intelligence transparency for end-users. Recently, interpretability has received considerable attention, especially since the European Parliament imposed the general data protection regulation (GDPR) in May 2018, which requires industries to "explain" any decision made when automated decision making occurs: "a right of explanation for all individuals to obtain meaningful explanations of the logic involved." The

current state of regulations is mainly focused on user data protection and privacy; it is expected to cover more algorithmic transparency and explanations requirements from artificial intelligence systems (Goodman & Flaxman, 2017). Additionally, the European Commission recently published a proposal for an Artificial Intelligence (AI) act that requires the development of trustworthy AI systems. The proposal clearly requires AI systems to make use of Explainable AI tools to increase transparency and interpretability (Act, 2021).

Addressing such a broad array of expectations for interpretability and transparency requires multi-disciplinary research efforts, as existing communities have different requirements and have other priorities and areas of specialization. For example, research in machine learning aims to design new interpretable frameworks and explain black-box models with ad-hoc explainers. Along the same line but with different techniques, visual analytics researchers study tools and data methods to enable domain experts to visualize complex black-box models and study interactions to manipulate machine learning models. In contrast, research in human-computer interaction (HCI) focuses on end-user needs such as user trust and understanding of machine-generated explanations. Psychology research also studies the fundamentals of human understanding, interpretability, and the structure of explanations.

Despite the increasing usage of machine learning-based prediction models in the medical sector, clinicians find it difficult to trust these models in practice (Darcy, Louie, & Roberts, 2016). Most of the models developed by data scientists primarily focus on prediction accuracy as a performance indicator, but they seldom explain their predictions in a meaningful way (Basu Roy et al., 2015). In addition to healthcare's ethical requirements and regulations, the lack of interpretability can result in life-threatening consequences. For example, Caruana et al. (Caruana et al., 2015) proposed a machine learning model for predicting the risk of readmission for patients with pneumonia. The model predicted that a patient had a lower risk of in-hospital death when admitted for pneumonia given asthma. Counterintuitively, patients with asthma are at higher risk of severe complications, including death, from an infectious pulmonary disease like pneumonia. In fact, the data was biased because these high-risk patients with Asthma were given special attention during their hospital visits which contributed to their lower mortality. The presence of Asthma was not responsible for their improvement in health, but rather a systematic bias. As a result, more concerns about interpretability, fairness and biases have been grown recently in the healthcare domain where human lives are at risk (Chen, Johansson, & Sontag, 2018).

In this work, we focus on techniques for extracting concepts from high-dimensional medical records of cardiorespiratory fitness. In these settings, the tabular raw data consists of numerous raw features. The clinician's mental model needs to comprehend these features and respond at a higher level of the patient condition (e.g. patient has an increased risk of obesity). Converting such low-level features into meaningful concepts that clinicians can readily reason about and then utilizing such concepts in explaining the prediction of an instance makes it easier to understand than providing an explanation in terms of low-level features. The current concept-based explanation techniques suffer from the following limitations that prevent their usage in the clinical setting: 1) the concepts are defined as a black-box model that may fail to capture the clinician's mental model, 2) these techniques assume the availability of ground-truth concept labels that may not be realistic in many application domains.

We summarize our contributions as follows:

- A novel local model-agnostic interpretability framework that provides concept-based explanation in the form of intuitive concepts deemed important to the prediction of the instance being explained.
- A counterfactual explanation, suggesting the minimum changes in the important concepts for the prediction of the instance being explained, led to a different outcome.

The remainder of this paper is organized as follows. We discuss the related work in Section 2. Section 3 describes the proposed concept-based interpretability framework. The results of our experiments are presented and discussed in Section 4 before we conclude the paper in Section 5.

2. Related Work

Our work relates to bias and fairness in Machine Learning and the interpretability of machine learning models.

2.1 Bias and Fairness in Machine Learning

While AI is promising to revolutionise medical practice, it faces substantial technological challenges. It's important to collect data representative of the target patient group. For example, data from various healthcare settings may cause a model trained on the data of one hospital to fail to generalise to another due to different forms of bias and noise in the data (Obermeyer & Emanuel, 2016). Krause et al. (Krause, Dasgupta, Swartz, Aphinyanaphongs, & Bertini, 2017) show how to detect biases in healthcare data using aggregated instance-level explanations. They used an instance-level algorithm optimized for sparse binary input data (Martens and Provost (Martens & Provost, 2014)). Through aggregation, filtering, and reordering, they discovered biases in their data for predicting hospital admission which made it impossible for the machine learning model to correctly predict admission in some cases. The model was aware that a CET or PET scan was taking place but was oblivious of the results. As a result, the model could not predict the diagnosis because the scan results directly impacted the outcome. Fairness evaluation and bias mitigation have been recently studied for tasks such as mortality prediction (Martinez, Bertran, & Sapiro, 2020; Zhang, Lu, Abdalla, McDermott, & Ghassemi, 2020; Chen, Szolovits, & Ghassemi, 2019), phenotyping (Zhang et al., 2020), readmission (Zhang et al., 2020), length of stay (Cui, Pan, Zhang, & Wang, 2020). It is also well acknowledged that enhancing model interpretability is an important step towards developing fairer ML systems (Doshi-Velez & Kim, 2017) since interpretations can help detect and mitigate bias during data collection or labeling (Lipton, 2018; Du, Yang, Zou, & Hu, 2020; Adebayo & Kagal, 2016).

2.2 Interpretability of Machine Learning Models

We have witnessed a notable explosion in the number of explanation techniques over the last few years due to the widespread need for explainable artificial intelligence (Bodria, Giannotti, Guidotti, Naretto, Pedreschi, & Rinzivillo, 2021). As a result, several recent studies

focused on the exploration of explainability in healthcare (Zhang, Xie, Xing, McGough, & Yang, 2017; Holzinger, Biemann, Pattichis, & Kell, 2017; Tonekaboni, Joshi, McCradden, & Goldenberg, 2019; Holzinger, Langs, Denk, Zatloukal, & Müller, 2019; Khodabandehloo, Riboni, & Alimohammadi, 2021). More specifically, specific analyses have been studied, e.g., chest radiography (Kallianos, Mongan, Antani, Henry, Taylor, Abuya, & Kohli, 2019), emotion analysis in medicine (Zucco, Liang, Di Fatta, & Cannataro, 2018), COVID-19 detection and classification (Lundberg, Erion, Chen, DeGrave, Prutkin, Nair, Katz, Himmelfarb, Bansal, & Lee, 2020), and the research encourages understanding of the importance of interpretability in the medical field (Langlotz, Allen, Erickson, Kalpathy-Cramer, Bigelow, Cook, Flanders, Lungren, Mendelson, Rudie, et al., 2019). The category of technical contribution in the topic of interpretability can be classified into two categories: *global* or *local* (Guidotti, Monreale, Ruggieri, Turini, Giannotti, & Pedreschi, 2018). In principle, global explanation techniques focus on the general prediction model decisions. In contrast, local explanation techniques focus on specifics of each instance and provide explanations that can lead to a better understanding of the features that contributed to the prediction of this instance based on smaller groups of instances that are often overlooked by the global interpretation techniques (Plumb, Molitor, & Talwalkar, 2018; Ribeiro, Singh, & Guestrin, 2016b; White & Garcez, 2019; ElShawi, Sherif, Al-Mallah, & Sakr, 2019; Panigutti, Guidotti, Monreale, & Pedreschi, 2019; Panigutti, Perotti, & Pedreschi, 2020). Another way to classify interpretability techniques is according to the problem they can solve (Mohseni, Zarei, & Ragan, 2021). Intrinsic interpretability is achieved by constructing self-explanatory models in which interpretability is directly inherited from their structures. The family of this category includes decision tree, rule-based model, linear model, etc. In contrast, the post-hoc interpretability requires creating a second model to provide explanations for an existing model. The main difference between these two categories lies in the trade-off between model performance and explanation fidelity. Inherently interpretable models could provide an accurate and undistorted explanation but may sacrifice prediction performance to some extent. The post-hoc ones are limited in their approximate nature while keeping the underlying model accuracy intact.

Since deep neural networks (DNN) achieved great success in different application domains (Domhan, Springenberg, & Hutter, 2015), there has been significant attention for developing various interpretability techniques for explaining such models. Most of the recent literature focused on visualizing and explaining the prediction of neural networks. Plenty of approaches developed to visualize the inner working of DNN. Visualizing the behaviour of DNN can be achieved by sampling patches that maximize activations of hidden units (Zeiler & Fergus, 2014), and by backpropagation to highlight the main features in the image that contributed to the prediction (Mahendran & Vedaldi, 2015; Simonyan, Vedaldi, & Zisserman, 2013; Zhou, Khosla, Lapedriza, Oliva, & Torralba, 2016; Selvaraju, Das, Vedantam, Cogswell, Parikh, & Batra, 2016). Such backpropagation-based techniques have been subsequently used for digital healthcare, especially in medical image analysis. Lee et al. (Lee, Yune, Mansouri, Kim, Tajmir, Guerrier, Ebert, Pomerantz, Romero, Kamalian, et al., 2019) developed an explainable technique for detecting acute intracranial haemorrhage from small datasets that is one of the most famous studies using Class activation mapping (CAM) (Zhou et al., 2016). Kim et al. (Kim, Kim, Kim, & Kim, 2021) summarised AI based breast ultrasonography analysis with CAM technique. In Hu et al. (Hu, Gao, Niu,

Jiang, Li, Xiao, Wang, Fang, Menpes-Smith, Xia, et al., 2020), a COVID-19 classification system was implemented with multiscale CAM to highlight the infected areas. The main drawback of such heatmap explanation techniques is that they are not informative enough to explain the main reasons for a particular prediction. Another line of research focused on explaining DNN by highlighting the most important features contributed to the prediction. Such techniques follow a common approach: for each input to the neural network model, change individual features either by removal (Michie, Spiegelhalter, Taylor, et al., 1994; Ribeiro, Singh, & Guestrin, 2016a) or perturbation (Sundararajan, Taly, & Yan, 2017; Ribeiro et al., 2016b; ElShawi et al., 2019) to approximate the contribution of each feature for the model's decision. Such "feature-based" explanation techniques suffered from several limitations (Ghorbani, Abid, & Zou, 2019a; Gimenez, Ghorbani, & Zou, 2018). Kindermans et al. (Kindermans, Hooker, Adebayo, Alber, Schütt, Dähne, Erhan, & Kim, 2019) showed that these techniques are vulnerable to simple shifts in the input. Human-based experiments showed that these techniques are prone to human biases and do not increase human trust in the black-box models (Kim, Wattenberg, Gilmer, Cai, Wexler, Viegas, & Sayres, 2017; Poursabzi-Sangdeh, Goldstein, Hofman, Vaughan, & Wallach, 2018). In addition, these experiments showed that given identical feature-based explanations, humans reached completely contradicting conclusions (Kim et al., 2017).

As a consequence, another line of research considered explaining predictions in the form of high-level understandable concepts (Kim et al., 2017; Zhou, Sun, Bau, & Torralba, 2018). Instead of giving the explanation in the form of an importance score for each input feature, the explanation reveals the main contributing concepts to the prediction. In (Feng, Min, Chen, Chen, Xie, Wang, & Chen, 2017), authors propose a multichannel convolutional neural network based on embeddings of medical concepts to examine the effect of patient characteristics on total hospital costs and length of stay. Mincu et al. (Mincu, Loreaux, Hou, Baur, Protsyuk, Seneviratne, Mottram, Tomasev, Karthikesalingam, & Schrouff, 2021) defined "clinical concepts" from temporal EHR input features to improve the human-understandability of post-hoc explanations of continuous clinical predictions. The main limitation of these methods is that they provide explanations based on the user's queries about concepts rather than considering the significant concepts for the prediction that users may not know about. More specifically, these methods require users to provide a set of hand-labelled examples for each concept of interest; the user needs to query its significance for the prediction, which could be challengeable. These methods are beneficial and provide great insights when the user knows exactly the set of concepts and has enough examples for each of these concepts. However, the space for meaningful concepts to be queried is unlimited, and in some cases, it is hard to provide enough examples for each of these concepts. Another primary limitation of these methods is that querying a particular set of concepts may create a biased explanation process toward such provided concepts while failing to query the right set of concepts.

As a result, a recent line of research has focused on automatic concept extraction. For example, Ghorbani et al. (Ghorbani, Wexler, Zou, & Kim, 2019b) developed a framework to automatically extract meaningful concepts that are important for the model prediction and then decompose the evidence for a prediction into an importance score for each of the extracted concepts. Another example, Elshawi et al. (El Shawi, Sherif, & Sakr, 2021) proposed a framework to automatically identify high-level human-understandable concepts

which are important for the convolution neural network for explaining the prediction of images by aggregating related local image segments (concepts) across diverse data and then decomposing the evidence for a prediction into such concepts through a shallow decision tree. While these methods allow the automatic extraction of meaningful concepts, they do not incorporate users feedback, making it hard for users to obtain concepts aligning with their intuitive perception of the problem. In contrast to this, our proposed approach learns concepts that align with users' knowledge about what a concept means and approximate the behaviour of the black-box model in the vicinity of the instance being explained through a fully transparent post-hoc model that does not require a dataset of instances labelled with concepts. Such learnt concepts map features that are difficult to interpret in high-dimensional domains to human-understandable concept representation.

2.3 Desired Characteristics for Local Concept-based Interpretability Techniques

Our goal is to provide explanations for the predictions of machine learning models in terms of units that are easier to comprehend by humans than individual low-level features. Following the literature (Zhou et al., 2018; Kim, Wattenberg, Gilmer, Cai, Wexler, Viegas, et al., 2018), in this work, these units are referred to as concepts. In the following, we outline a number of desired characteristics that should be satisfied by local concept-based explanation techniques.

1. Meaningfulness: a concept has meaning for users by its own. In the case of tabular data, for example, individual features may not meet this criteria, yet a collection of features remains meaningful. Meaningfulness should also correspond to the alignment with human' knowledge about what a concept means.

2. Model-agnostic: a concept-based explanation technique should be able to explain *any* model. Apart from the fact the many state-of-the-art machine learning models are not currently interpretable, this also provides flexibility to explain future machine learning models.

3. Counterfactual actions: a concept-based explanation should provide minimum changes necessary to change the prediction of the instance being explained.

4. Local fidelity: local explanation methods should approximate the behaviour of the back-box model in the vicinity of the instance being explained.

5. Interpretable: concept-based explanation techniques should provide explanations that are comprehensible by humans. Thus, linear models may not be interpretable if hundreds of features significantly contribute to a prediction. It is not reasonable to expect a user to comprehend an explanation by inspecting the weights of hundred features. This requirement further implies that explanations should be easy to understand, and thus the "low-level input" features may need to be different from the ones used in the explanation.

There is no broad agreement upon the properties that should be satisfied by the local concept-based explanations; however, we believe that meeting these properties is a good starting point toward intuitive concept-based explanations.

3. Framework for Local Model Agnostic Concept-based Interpretability

The process of explaining individual predictions is illustrated in Figure 1. It is clear that a clinician is much better positioned to make a decision with the help of a machine learning model if meaningful explanations that align with his/her knowledge are provided (Ribeiro et al., 2016b). In this case, an explanation is a small set of concepts contributing to the prediction of the instance being explained. Clinicians have acknowledged that providing explanations in the form of concepts increases their trust in the black-box machine learning model used.

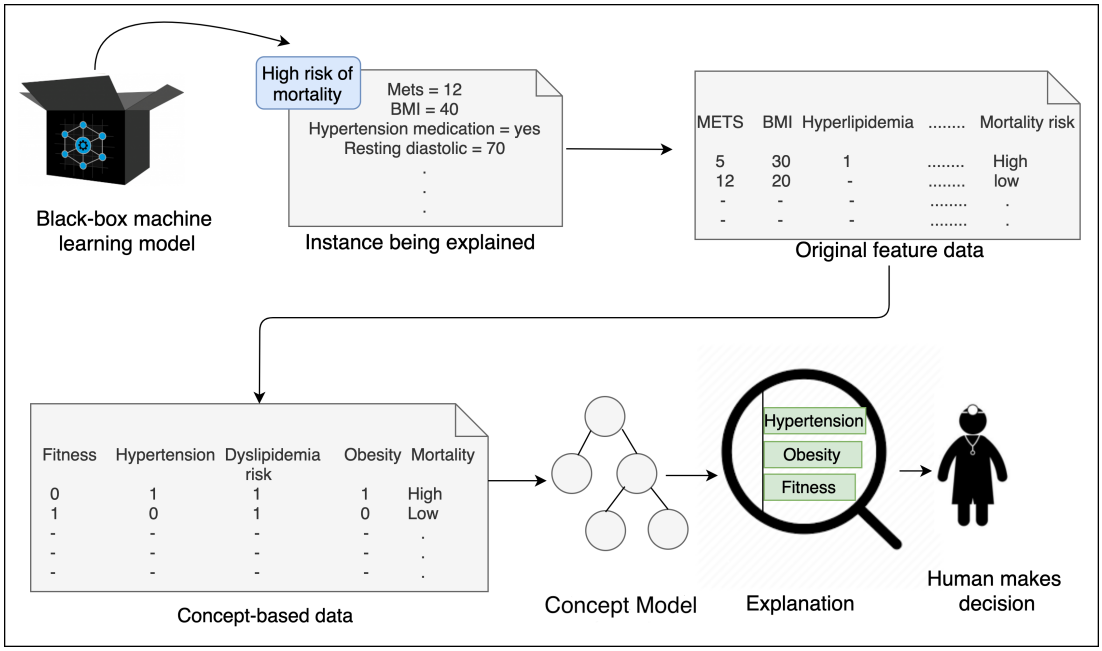

Figure 1: Explaining individual predictions. The patient being explained is represented in terms of low-level features including vital signs, diagnosis and clinical laboratory measurements. A black-box model predicts this patient as high risk of mortality. CLEF maps the input representation (patient's history data) to an intermediate concept-based representation that uses high-level intuitive concepts. Next, CLEF learns a model (concept model) on such concepts to decompose the evidence of the prediction of the instance being explained into high-level intuitive concepts. Concepts hypertension, obesity, and fitness are portrayed as contributing to the "high risk of mortality" prediction. With these, a doctor can make an informed decision about whether to trust the model's prediction.

In the following, we present CLEF, a novel local model-agnostic framework for learning a set of high-level transparent concept definitions in high-dimensional tabular data. The overall goal of CLEF is to learn a local model over concepts that align with clinician knowledge while being locally faithful to the black-box model.

3.1 Fidelity-Interpretability Trade-off

We denote $x' \in R^d$ be the original representation of an instance being explained. Formally, we define an explanation as a model $f \in F$ built on the top of high-level intuitive concepts, where F is a class of potentially transparent models, such as linear models and decision trees. Let the model being explained be denoted z. In classification, $z(x')$ is the probability (or a binary indicator) that x' belongs to a certain class. We further use $\pi_{x'}(t)$ as a proximity measure between an instance t to x', so as to define locality around x'. Finally, let $L(z, f, \pi_{x'})$ be a measure of how unfaithful f is in approximating the behaviour of z in the locality defined by $\pi_{x'}$. Let Ω be a measure for how complex the explanation model f. For example, for a linear model, $\Omega(f)$ may be the number of non-zero weights. To satisfy both interpretability and local fidelity properties, we must minimize $L(z, f, \pi_{x'})$ while having $\Omega(f)$ low enough to be interpretable by humans. The explanation produced by CLEF is obtained by the following:

$$\zeta(x') = \arg\min_{f \in F} L(z, f, \pi_{x'}) + \Omega(f) \tag{1}$$

Given a training dataset $\{x_n, y_n\}^N$, we aim to learn a 2-stage prediction function f that approximates the behaviour of z in the vicinity of x', where x is the input feature vector and $y \in \{0, 1\}$ is the prediction of z. The first function, denoted concept definition g, maps the low level features x to concepts $c \in \{0, 1\}^C$. The second function, f, maps concepts c to y. Our goal is to learn f that is interpretable and locally faithful to z, while learning g that is intuitive in a way that models clinician knowledge.

3.2 Sampling for Local Exploration

Our goal is to minimize the locality-aware loss $L(z, f, \pi_{x'})$ as in equation 1 without making any assumption about z, since we want CLEF to be model-agnostic. To capture the behaviour of z in the vicinity of x', we approximate $L(z, f, \pi_{x'})$ by drawing samples weighted by $\pi_{x'}$. More specifically, we randomly sample a set of instances $S_{x'}$ from $\{x_n, y_n\}^N$ and weight sample instances by their proximity from x' such that sample instances in the vicinity of x' are assigned a high weight, and far away instances from x' are assigned low weight. In this work, the size of a sample $S_{x'}$ is chosen to be 1000, leaving the exploration of dynamic sample size for future work. Given the dataset $S_{x'}$, we optimize equation 1 to get explanation $\zeta(x')$. CLEF presents an explanation that is locally faithful, where the locality is captured by $\pi_{x'}$.

3.3 Learning Interpretable Concepts with Human Feedback

In the following, we show how to learn functions g and f such that g is intuitive and closely align with clinician knowledge about concept definition, while f is faithful in approximating the behaviour of z. Our definition for the concepts is inspired by the medical literature, where conditions are usually defined from high dimensional medical records. Such form of concept definition is interpretable for clinicians as it is the defacto clinical technique (Castro, Minnier, Murphy, Kohane, Churchill, Gainer, Cai, Hoffnagle, Dai, Block, et al., 2015). We define a binary matrix $A \in \{0, 1\}^{D \times C}$, where D is the number of features of the dataset $S_{x'}$, $A_{i,j} = 1$ represents the association of feature x_i to concept j and $A_{i,j} = 0$ represents the

dissociation of feature x_i from concept c_j. A concept c exists in a particular instance x if at least one of the features associated with c exists in x. The main goal of this approach is to learn the set of features associated with each concept. The decomposition of the prediction of the instance being explained into concepts enables an interpretable explanation of the prediction. Since our goal is to interpret the instance being explained in terms of the high-level concepts rather than raw input features, the prediction function f that is dependent on the concepts should be interpretable.

To ensure the meaningfulness of the explanations provided by CLEF, we learn intuitive concepts that align with clinician knowledge while incorporating clinicians' feedback into the learning process. More specifically, the clinician is expressly asked if a feature x_i should be connected with a concept c_j. For example, an association between the feature 'insulin' and the concept' diabetes' might make sense, whereas an association between 'insulin' and 'hypertension' does not make sense, even though it might make the concept more predictive. Our definition of intuitiveness is inspired by (Lage & Doshi-Velez, 2020), where the intuitiveness of function g is satisfied if the user accepts the suggested association between a particular feature x_i and a concept c_j for every (i, j) feature-concept association in g. To learn g that satisfies intuitiveness, we do the following. First, initialize matrix, A, by asking clinicians to specify one feature they wish to associate to each concept. Clinicians are usually familiar with high-level concepts of interest that affect the prediction of the risk of mortality, in addition to few features associated with each concept, but it is hard to come up with the long tail of the features related to each of the concepts; this is where the proposed technique is useful. We summarize the process of associating features to concepts in Algorithm 1. The algorithm builds up g on $S_{x'}$ iteratively by making a number of feature-concept (i^*, j^*) proposals that clinician either accept or reject. Such proposals are made from pairs of (i, j) that the algorithm has not yet explored. For each concept, we make a fixed number of proposals before moving to the next concept. In this work, we use a fixed number of proposals per concept $numproposals = 7$. More specifically, each concept c_j is associated with two list of features; the explored list l_j consists of features that have been proposed to a clinician to be associated with concept c_j and the other list u_j consists of the set of features that have not been proposed yet for concept c_j. If the clinician accepts the proposed feature-concept association, then the proposed feature is added to the concept definition and thus feature-concept matrix $A_{i,j} = 1$; otherwise, the feature-concept matrix remains unchanged. List l_j is first initialized with a single feature i, such that $A_{i,j} = 1$ for each concept j and u_j is initialized with the rest of features that are not included in l_j. Algorithm 1 models the human feedback while proposing feature-concept associations by incorporating the clinician's prior acceptance of feature-concept associations to improve future proposals made by the algorithm and refit model f each time g is updated. To do so, we store a set of labels of the proposals that the user has previously accepted or rejected in matrix $intuit$. This matrix is first initialized so that $intuit_{i,j} = 1$ and $intuit_{i,j' \neq j} = 0$ if $A_{i,j} = 1$ in the concept definitions initialized by the user. The matrix is then updated such that $intuit_{i^*,j^*} = 1$ if the user accepts the proposed feature-concept association; otherwise, it remains unchanged. We assume that a single feature can be associated with different concepts.

The key challenge is to propose feature-concept associations that are intuitive for the clinician and equally highly faithful to the model being explained. If the proposal is highly

Algorithm 1: Algorithm for interactively proposing intuitive and interpretable concepts with human feedback

Input : $S_{x'}$, A, $numproposals$
Initialize: l, u, $intuit$
$J^* \leftarrow 1$
while $J^* \leq numconcepts$ **do**
  $k \leftarrow 1$
  while $k \leq numproposals$ **do**
    Calculate $SFid_{i^*,j^*}$ for all instances in u_{j^*}
    Calculate $SIntuit_{i^*,j^*}$ for all instances in u_{j^*}
    Select the best feature i^*, by constructing a pareto-front based on the trade-off between $SFid$ and $SIntuit$
    if (i^*, j^*) *is accepted* **then**
      $intuit_{i^*,j^*} = 1$
      $A_{i^*,j^*} = 1$
      Retrain f
    else
      $intuit_{i^*,j^*} = 0$
    end
    $l_{j^*} = l_{j^*} \cup \{i^*\}$
    $u_{j^*} = u_{j^*} \backslash \{i^*\}$
    $k \leftarrow k + 1$
  end
  $J^* \leftarrow J^* + 1$
end

faithful to the black-box model but not intuitive, then the clinician will not accept it, and no improvement will be achieved in f. On the other side, if the proposal is unfaithful, then it will not improve f even if the clinician accepts the proposal. The goal is to make a reasonable number of proposals that are both intuitive and highly faithful. To achieve this target, we compute two scores for fidelity $SFid$ and intuitiveness $SIntuit$ for each proposal. The goal of $SFid_{i,j}$ is to measure how well our model f is capturing the behaviour of z in the vicinity of x' when associating feature i to concept j. For each concept c_j, we calculate $SFid_{i,j}$ by updating f if the proposal (i, j) is accepted by the clinician. The goal of $SIntuit_{i,j}$ is to assess the likelihood of the acceptance of the association of feature i to concept j by the clinician. For each concept c_j, we calculate $SIntuit_{i,j}$. We assume that a clinician will likely accept a proposal that associates a feature i to a concept j if a feature i' similar to i has been associated before to concept j. The notion of similarity between two features is defined by the Jaccard similarity (denoted J) computed over the number of times each feature is recorded for each instance (x^T). The probability that a clinician will accept associating feature i to concept j is calculated through similarity graph as follows:

$$SIntuit_{i,j} = exp(\frac{1}{2}\Sigma_{i' \in l_j} J(x_i^T, x_{i'}^T)(Intuit_{i,j} - Intuit_{i',j})^2) \quad (2)$$

To make feature-concept proposals that are highly faithful and intuitive, we rank proposals based on the Pareto front of the trade-off between the intuitiveness and fidelity. The proposal with the highest rank from the Pareto front is selected.

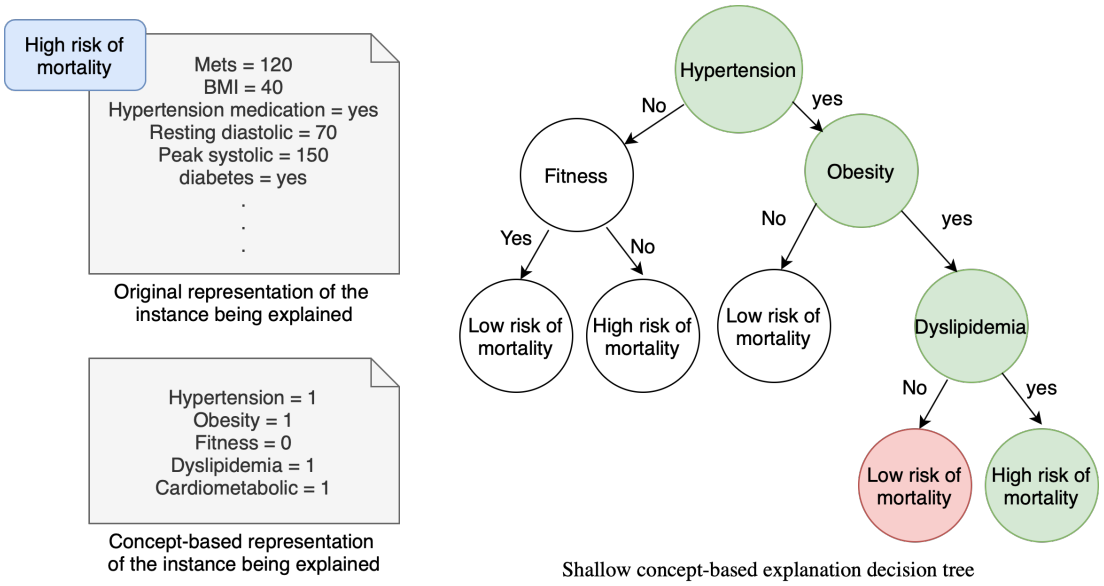

Figure 2: Shallow concept-based explanation decision tree of depth 4 explaining the prediction of a patient of high risk of mortality

3.3.1 Constructing Local Explanation

The CLEF is based on the view that a satisfactory explanation of a single prediction needs to explain the value of that prediction and answer' what-if-things-had-been-different' questions. The CLEF framework considers two different explanation models to provide counterfactual explanations. The first explanation model is a decision tree classifier. It is used due to its interpretable nature that allows concept rules to be derived from a root-leaf path in the decision tree and counterfactuals that can be extracted by symbolic reasoning over a decision tree. To guarantee a fast and easy search for counterfactuals, we consider all possible paths in the decision tree leading to a decision that is not equal to the decision of the instance being explained x'. Among all these paths, we only consider the one with the minimum number of spilt conditions that are not satisfied by instance x'. Increasing the depth of a decision tree increases the prediction accuracy, which leads to less interpretable results as the number of nodes increases exponentially with depth. Thus, a shallow decision tree is favourable as it is more comprehensible by humans. In this work, we use a fixed depth of 4, leaving the exploration of dynamic depth to future work. Figure 2 shows a decision tree explanation of a patient of high risk of mortality. It is clear from the explanation tree that the patient has been predicted at high risk of mortality because of the existence of concepts 'hypertension', 'obesity', and 'dyslipidemia'. As a further output, CLEF computes

a counterfactual explanation which is the path in the decision tree corresponding to the existence of concepts 'hypertension', 'obesity', and the absence of concept 'dyslipidemia'that leads to the prediction of the instance being explained as low risk of mortality. The second explanation model is logistic regression due to its interpretable nature that allows concepts to be explained through their weights. To generate a counterfactual explanation from the logistic regression model, we do the following. Let x'' be the representation of the instance being explained x' in terms of high level concepts learnt in Section 3.3. Let $min_c(x'')$ denote a vector resulting from changing the value of one concept c in x'' such that $f(min_c(x''))$ $= y'$ and $f(x') = y$, where $y \neq y'$. A perturbation of x' is defined as the change in the value of concept c to flip the prediction of x'. We compute all the perturbations of x' for all concepts and finally returns the perturbation with the highest probability of class y'.

4. Results and Discussion

In this section, we introduce the dataset used in this work and the concepts definition in Sections 4.1, and 4.2, respectively. We define baselines in Section 4.3 to be compared to the proposed approach in Section 4.4. The faithfulness of the proposed approach is evaluated in Section 4.5. We evaluate the trust in the explanations of CLEF in Section 4.6. We show in Section 4.7 the effectiveness of the explanations of CLEF in detecting bias in data.

4.1 Henry Ford FIT Dataset

The dataset of this study was collected from patients who underwent treadmill stress testing by physician referrals at Henry Ford Affiliated Hospitals in metropolitan Detroit, MI in the United States, FIT Project (Al-Mallah et al., 2014). In particular, the data obtained from the electronic medical records, administrative databases, and the linked claim files and death registry of the hospital over the period between 1 January 1991 and 28 May 2009. Study participants underwent routine clinical treadmill exercise stress testing using the standard Bruce protocol between 1 January 1991 and 28 May 2009. The dataset includes 43 attributes containing information on vital signs, diagnosis and clinical laboratory measurements. Examples of these attributes include sex, age, heart rate achieved, resting systolic blood pressure, resting diastolic blood pressure, obesity, hypertension, diabetes, and METS. In this study, we have excluded from the original registry of the FIT project the patients with known coronary artery disease ($n = 10,190$) or heart failure ($n = 1162$) at the time of the exercise test or with less than a 10-year follow-up ($n = 22,890$). Therefore, a total of 34,212 patients were included in this study. After a 10-year follow-up, a total of 3,921 patients (11.5%) died, as verified by the national social security death index. All included patients had a social security number. In this work, we classified patients into two categories: low risk of all-cause mortality (ACM) and high risk of ACM. In particular, patients were considered to have a high risk for ACM if the predicted event rate is more than or equal to 3%.

4.2 Concepts Definition

The dataset used in this work is split 60% for training, 20% for validation and 20% for testing. To quantitatively evaluate the proposed approach and compare it to multiple

baselines, we ran experiments with known handcrafted concepts defined by a clinician to be discovered from real data. We seeded each experiment with features from known concepts, and we assumed that the proposal of features belonging to these concepts is accepted by the user. We relied on clinicians to define a set of handcrafted concepts and associate the ground truth features to each of the concepts. The list of features associated with each concept was compiled by the second author. The concepts are defined as follows 'Fitness', 'Hypertension', 'Obesity/diabetes', 'Dyslipidemia', and 'Cardiometabolic'. The associated features for each concept are defined as follows

- 'Fitness': mets_achieved > 10, peak systolic blood pressure > 200
- 'Hypertension': hypertension = yes, hypertension medication = yes, calcium channel blockers = yes, diuretics = yes, angiotens in receptor blocker = yes, angiotensin–converting enzyme inhibitor = yes, beta blockers = yes
- 'Obesity/diabetes': body mass index > 30, diabetes = yes, diabetes medication = yes, insulin = yes, glycated hemoglobin > 7
- 'Dyslipidemia': body mass index > 30, statin use = yes, hyperlipidemia = yes, hyperlipid = yes, hyperlipidemia = yes, low-density lipoprotein > 160, high-density lipoprotein < 40, chol > 200, triglyceride > 200
- 'Cardiometabolic': body mass index > 30, concept 3 ('Obesity/Diabetes') features, concept 4 'Dyslipidemia'features.

4.3 Baselines

To explain individual prediction, we compare our proposed approach CLEF to two baselines. The first one is an interactive concept-based baseline, and the other one is non-interactive. The interactive baseline is compared to our g (concept definitions) and has the same explanation function f trained on the top of concepts. For the interactive baseline, we need to simulate the clinician interaction of the baseline, which is equivalent to user feedback on feature-concept association in our approach. More specifically, the interactive baseline, denoted AL, fits five concept-classifier models (regularized logistic regression models), one for each concept. More specifically, for each instance x' in the testing dataset, we train a concept classifier for each concept c on a subset D_c of $S_{x'}$. Such subset is a mix of instances balancing the presence and absence of concept c. We define $D_c = D_c^+ \cup D_c^-$, where $D_c^+ = \{(x_1, y_{c1}), ..., (x_q, y_{cq})|_{y_c=1}\}$ and $D_c^- = \{(x_1, y_{c1}), ..., (x_q, y_{cq})|_{y_c=0}\}$, where $y_c \in \{0, 1\}$ indicates the absence or the presence of concept c in an instance x, and q is the number of examples in each of D_c^+ and D_c^-. Negative examples D_c^- for each concept c are selected randomly from other instances that do not have concept c such that the number of examples in D_c^+ and D_c^- are equal. We use these concept classifiers for each instance $x \in S(x')$ to create a vector $x_{AL} = (r_1, r_2, ..., r_5)$ representing the probability of each concept c in x. Next, we use concept vectors for instances in $S(x')$ directly in training unregularized logistic regression and shallow decision tree. The user's feedback is represented in labelling instances with concept labels. The non-interactive baseline do not employ concepts. Simply,

we compare to regularized logistic regression (LR). We train all approaches using the scikit-learn implementations (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, et al., 2011).

4.4 Comparison to Baselines

As a black-box model to be explained, we train a random forest model on the training dataset, and for each instance x' in the testing dataset, we report the performance of CLEF and all baselines on $S_{x'}$ sampled from the training dataset with class labels obtained from the random forest model. For more details about the random forest model for predicting the risk of mortality, we refer the readers to (Sakr, Elshawi, Ahmed, Qureshi, Brawner, Keteyian, Blaha, & Al-Mallah, 2017). The mean accuracy of predicting the risk of mortality (downstream accuracy) and the accuracy of mapping low-level features to concepts (concept accuracy) on the testing dataset for our approach and the baselines are reported in Table 1. The results show that our proposed approach, when f is either decision tree or logistic regression, outperforms the AL baseline on concept accuracy and downstream accuracy. Our final concept accuracy, when f is logistic regression, is 98%±0.002, which is 13% greater than the AL baseline. This substantial difference suggests that our proposed approach aligns much better with clinician intuitive representation than the baseline. Our approach with the two variants of f (logistic regression and decision tree) outperform the LR baseline by around 21%. Such baseline is equivalent to training f on original raw features of instances in $S_{x'}$ for each instance x' in the testing dataset. Such results suggest that training a decision tree or a logistic regression on top of the high-level concepts improves the predictive performance over training LR on the original raw features. In addition, our approach has a competitive advantage over the LR baseline, which is the predictors used in our approach are specified by the clinicians, whereas the inputs to LR do not have any constraint on their intuitiveness and colinearity, while concepts are guaranteed to represent different aspects.

Table 1: Downstream accuracies and concept accuracies on the testing dataset ± standard deviation for our proposed technique and baseline.

Variant	**Downstream accuracy**	**Concept accuracy**
Proposed approach when f is logistic regression	87% ±0.001	98% ±0.002
Proposed approach when f is decision tree	88% ±0.001	98% ±0.002
AL	77% ±0.001	85% ±0.003
logistic Regression (LR)	67% ±0.00	-

In Figure 3, we compare the downstream mean accuracy of our approach using the decision tree variant against randomly selected features from concept definitions to stimulate a user manually generating g. The x-axis represents the number of feature-concept proposals per concept. The reported results for our approach shown in Figure 3 are based on 7 proposals used to generate the results in Table 1, 5 proposals and 4 proposals. The

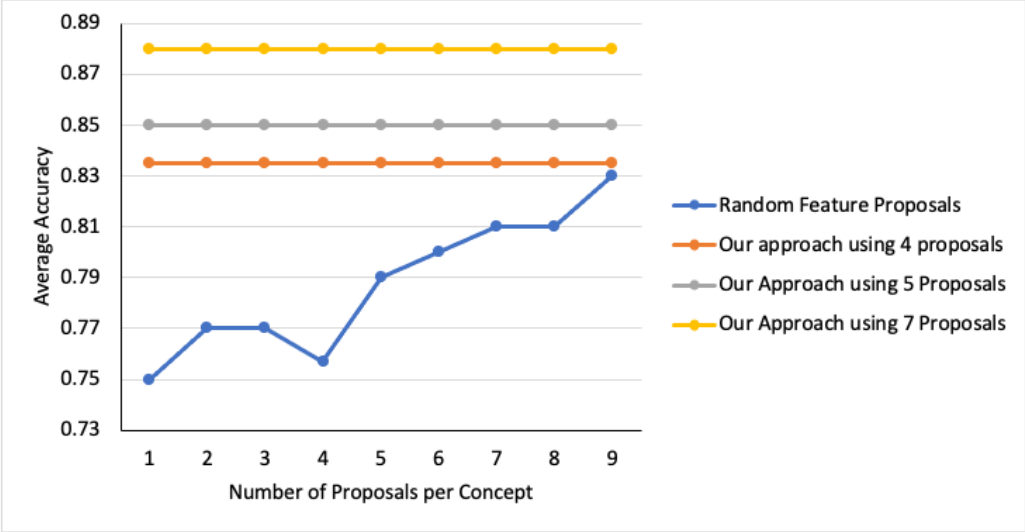

Figure 3: Downstream accuracy of our approach using decision tree variant against randomly selected features from the concept definitions

results show that adding random features from the concept definitions never approach the performance of our proposed approach, as shown in Figure 3

4.5 How Faithful are the Explanations of CLEF

In this experiment, we measure the faithfulness of the explanations obtained from CLEF on a classifier that is interpretable by its nature (sparse logistic regression). In particular, we train logistic regression classifier such that the maximum number of features used by any instance is 10, and thus, we know the *gold set* of features that the model considers. For each instance in the testing dataset, we generate the explanation from CLEF and AL. For each explanation retrieved by CLEF and AL, we compute the fraction of features returned in the top 3 concepts contained in the gold set. We report this recall averaged over all the instances in the testing dataset. The results show that CLEF achieves an average recall of 93% while AL achieves 89%, demonstrating that CLEF explanations are faithfull to the model being explained.

4.6 Can We Trust the Explanations of CLEF?

In this experiment, we measure the quality of the explanations of CLEF and measure how trusted the explanations obtained from CLEF are compared to AL. First, we train two black-box models (random forest and support vector machine) on the training dataset and get the prediction of each instance x' in the testing dataset from each black-box model. Next, we randomly select 20% of the features of instance x' and create a new instance w' with the same feature values of x' but with random values for the randomly selected features. Then, we get the prediction of instance w' from each black-box model. We assume that we have a trustworthiness oracle that labels a test instance as trusted if the prediction of

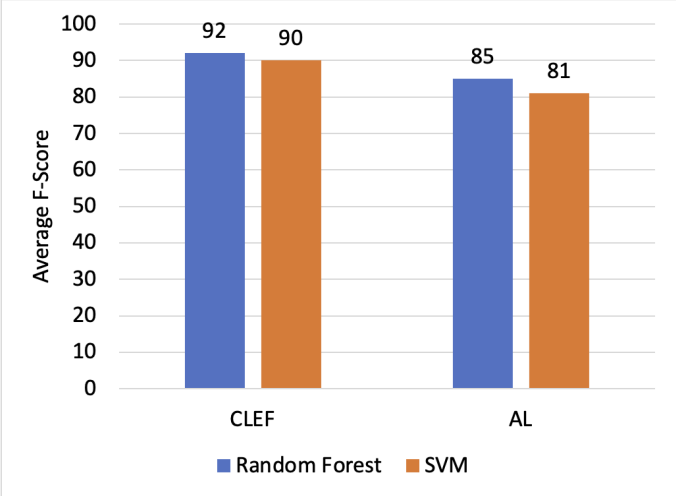

Figure 4: Average F1 of trustworthiness for CLEF and AL averaged over 50 runs on different classifiers (random forest and support vector machine).

w' from the black-box model equals the prediction of x' and untrusted otherwise. For each instance x' in the testing dataset, we get the explanation of x' from CLEF and AL. Let w'' be the representation of w' in terms of high level concepts learnt in Section 3.3 and let w'_{AL} be the concept vector of w', representing the probability of each of the concept classifiers trained for the interactive baseline AL in Section 4.3. We get $f(w'')$, and $f(w'_{AL})$. An instance x' is trusted if $f(w'') = f(x')$ for CLEF, and $f(w'_{AL}) = f(x')$ for AL baseline. We compare the trusted and untrusted instances for CLEF and AL against the trustworthiness oracle. Using this set-up, we report the overall F-score of CLEF and AL averaged over 50 runs using different black-box models, as shown in Figure 4. The results show that CLEF outperforms AL on the two black-box models.

4.7 Effectiveness of the Explanation of CLEF in Detecting Biases in the Data

In order to see whether the explanations obtained from CLEF are helpful in detecting biases in the training data, we created a modified version of the dataset used in this work with an inherent bias. We used visual analytics method to detect the bias on the dataset inspired by (Josua Krause, 2018). In particular, we compare the CLEF when f is logistic regression and LR baseline through an interface to detect the bias. We created a biased dataset such that diabetes, diabetes medication, and insulin features are inversely related to the risk of mortality; if the patient has diabetes, takes diabetes medication and insulin, then the patient is at low risk of mortality which is counterintuitive. We trained random forest on the biased and the unbiased datasets. The bias is created such that the biased model achieves higher testing accuracy than the unbiased model. The bias is created with the same degree in both training and testing datasets. The testing accuracies on the unbiased and biased datasets are 90% and 93%, respectively. The users who evaluated the bias were people with basic knowledge in the medical domain. For each of the biased and unbiased testing datasets, we

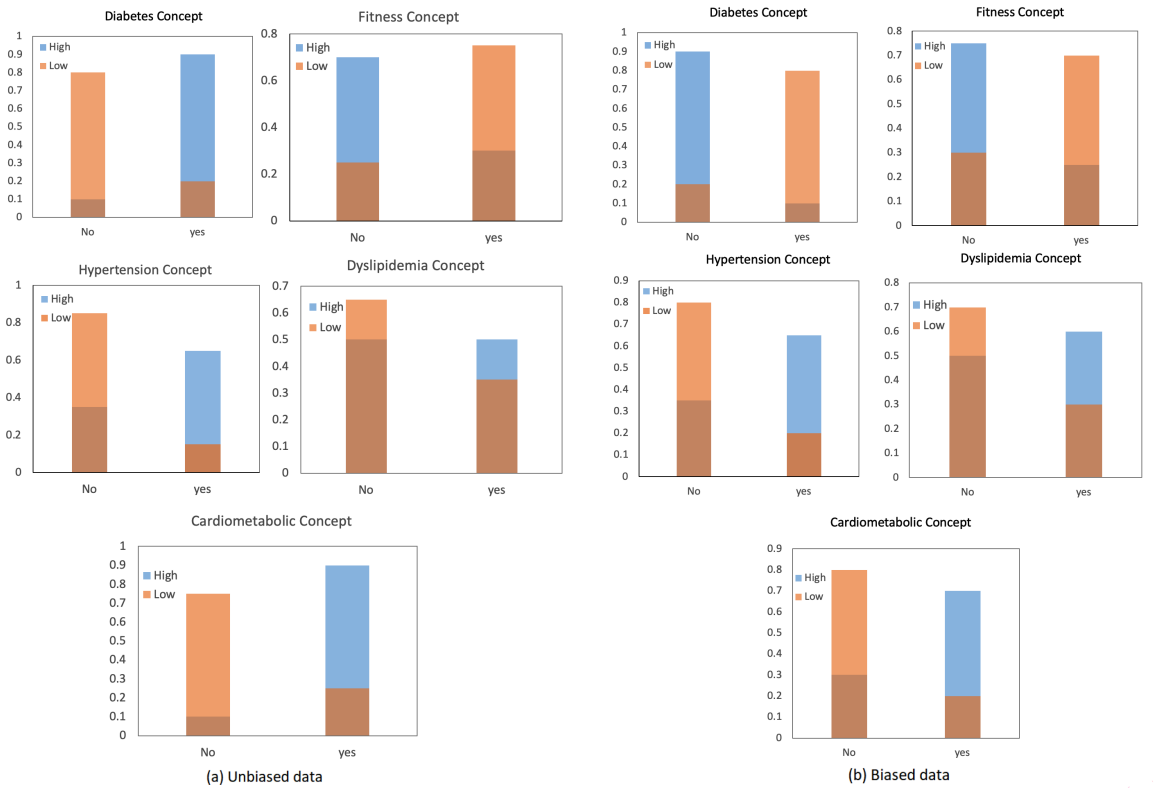

Figure 5: Explanation user interface on both the unbiased model (left side) and the biased model (right side) using CLEF

explain each instance from CLEF and the LR baseline. For the LR baseline, the explanation is in the form of the most important nine features; where importance is captured through weights. For comparing CLEF and LR for bias evaluation, we compare patients at high risk of mortality and patients at low risk of mortality. In particular, we show an aggregate user interface that shows the distribution of features/concepts values as histograms sorted such that the top-left histogram is for the feature/concept with the highest average contribution. The bottom-right histogram is for the feature/concept with the lowest average weight. The user interfaces for CLEF and LR, shown in Figure 6, illustrate 9 features for LR baseline and 5 concepts for CLEF. For each histogram, the height of the bars represents the percentage of instances in each group.

We conducted a user study to evaluate the ability to detect biases is the data by comparing the explanations of the biased and unbiased instances in the testing datasets using CLEF and LR baseline. This study involved 30 post-graduate students. We introduced the meaning of accuracy and how it is used to evaluate the model's performance. We then explained the mortality dataset by informing the participants of the meaning of the features and how they logically affect the output class, i.e. smoker patients are at higher risk of mortality than non-smokers. Finally, we explained to the participants how to use the evaluation interface. Out of the 30 participants, only 25 responses were valid. We evaluated validity

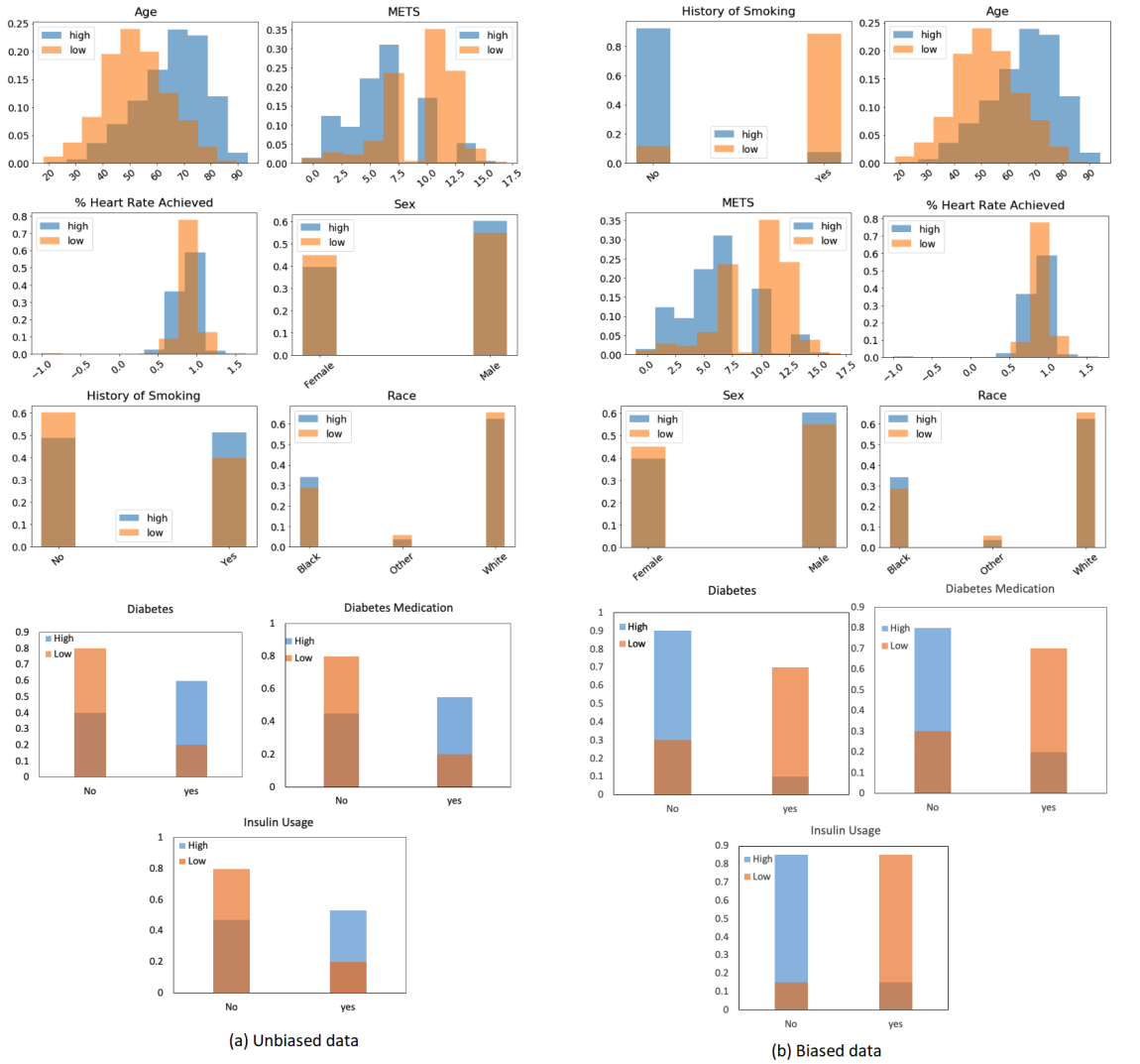

Figure 6: Explanation user interface on both the unbiased model (left side) and the biased model (right side) using LR baseline

by asking the participants an obvious question i.e., which model has better accuracy? All the participants were able to identify the bias from the CLEF interface, while only 80% were able to do the same using the interface of AL baseline. It is clear from the results that the explanation of CLEF enables participants to detect the bias better, demonstrating that CLEF explanations have a significant impact on users' ability to detect biases in data.

5. Conclusion

In this work, we proposed an interactive approach for locally explaining the risk of mortality in terms of high-level concepts. In addition, we provided counter-factual explanations that

explain the minimum number of concepts to be changed to flip the prediction. We used a human-in-the-loop approach for training concepts that align with clinician knowledge and we showed on a dataset that has been collected from patients who underwent treadmill stress testing with simulated concept definitions that our approach can learn representations that align with clinician intuitive concepts. The results show that the proposed approach provide explanations that are faithful to the model being explained. In addition, the explanations provided by CLEF have a significant impact on users' ability to detect biases in data. The proposed approach can be easily transferred to other domains. However, it has some limitations. In some domains, it is very difficult for users to identify concepts and associate features to these concepts. While our approach outperforms interpretable baseline, we did not test our approach extensively through human-based experiments.

Acknowledgments

The work of Radwa El Shawi is funded by the European Regional Development Funds via the Mobilitas Plus programme (grant MOBTT75).

References

Act, A. I. (2021). Proposal for a regulation of the european parliament and the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. *EUR-Lex-52021PC0206*.

Adebayo, J., & Kagal, L. (2016). Iterative orthogonal feature projection for diagnosing bias in black-box models. *arXiv preprint arXiv:1611.04967*.

Al-Mallah, M. H., et al. (2014). Rationale and design of the Henry Ford Exercise Testing Project (the FIT project). *Clinical cardiology*, *37*(8).

Basu Roy, S., et al. (2015). Dynamic hierarchical classification for patient risk-of-readmission. In *KDD*.

Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., & Rinzivillo, S. (2021). Benchmarking and survey of explanation methods for black box models. *arXiv preprint arXiv:2102.13076*.

Caruana, R., et al. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *KDD*.

Castro, V. M., Minnier, J., Murphy, S. N., Kohane, I., Churchill, S. E., Gainer, V., Cai, T., Hoffnagle, A. G., Dai, Y., Block, S., et al. (2015). Validation of electronic health record phenotyping of bipolar disorder cases and controls. *American Journal of Psychiatry*, *172*(4), 363–372.

Chen, I., Johansson, F. D., & Sontag, D. (2018). Why is my classifier discriminatory?. *Advances in Neural Information Processing Systems*, *31*.

Chen, I. Y., Szolovits, P., & Ghassemi, M. (2019). Can ai help reduce disparities in general medical and mental health care?. *AMA journal of ethics*, *21*(2), 167–179.

Cui, S., Pan, W., Zhang, C., & Wang, F. (2020). xorder: A model agnostic post-processing framework for achieving ranking fairness while maintaining algorithm utility.. *arXiv preprint arXiv:2006.08267*.

Darcy, A. M., Louie, A. K., & Roberts, L. W. (2016). Machine learning and the profession of medicine. *Jama*, *315*(6).

Domhan, T., Springenberg, J. T., & Hutter, F. (2015). Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Doshi-Velez, F., & Kim, B. (2017). A roadmap for a rigorous science of interpretability. *arXiv preprint arXiv:1702.08608*, *2*, 1.

Du, M., Yang, F., Zou, N., & Hu, X. (2020). Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, *36*(4), 25–34.

El Shawi, R., Sherif, Y., & Sakr, S. (2021). Towards automated concept-based decision treeexplanations for cnns.. In *EDBT*, pp. 379–384.

ElShawi, R., Sherif, Y., Al-Mallah, M., & Sakr, S. (2019). Ilime: Local and global interpretable model-agnostic explainer of black-box decision. In *European Conference on Advances in Databases and Information Systems*, pp. 53–68. Springer.

Feng, Y., Min, X., Chen, N., Chen, H., Xie, X., Wang, H., & Chen, T. (2017). Patient outcome prediction via convolutional neural networks based on multi-granularity medical concept embedding. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 770–777. IEEE.

Ghorbani, A., Abid, A., & Zou, J. (2019a). Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 3681–3688.

Ghorbani, A., Wexler, J., Zou, J., & Kim, B. (2019b). Towards automatic concept-based explanations. *arXiv preprint arXiv:1902.03129*.

Gimenez, J. R., Ghorbani, A., & Zou, J. (2018). Knockoffs for the mass: new feature importance statistics with false discovery guarantees. *arXiv preprint arXiv:1807.06214*.

Goodman, B., & Flaxman, S. (2017). European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, *38*(3), 50–57.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 93.

Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable ai systems for the medical domain?. *arXiv preprint arXiv:1712.09923*.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(4), e1312.

Hu, S., Gao, Y., Niu, Z., Jiang, Y., Li, L., Xiao, X., Wang, M., Fang, E. F., Menpes-Smith, W., Xia, J., et al. (2020). Weakly supervised deep learning for covid-19 infection detection and classification from ct images. *IEEE Access*, *8*, 118869–118883.

Josua Krause, Adam Perer, E. B. (2018). A user study on the effect of aggregating explanations for interpreting machine learning models. *KDD Workshops*.

Kallianos, K., Mongan, J., Antani, S., Henry, T., Taylor, A., Abuya, J., & Kohli, M. (2019). How far have we come? artificial intelligence for chest radiograph interpretation. *Clinical radiology*, *74*(5), 338–345.

Khodabandehloo, E., Riboni, D., & Alimohammadi, A. (2021). Healthxai: Collaborative and explainable ai for supporting early diagnosis of cognitive decline. *Future Generation Computer Systems*, *116*, 168–189.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2017). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). *arXiv preprint arXiv:1711.11279*.

Kim, J., Kim, H. J., Kim, C., & Kim, W. H. (2021). Artificial intelligence in breast ultrasonography. *Ultrasonography*, *40*(2), 183.

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., & Kim, B. (2019). The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 267–280. Springer.

Krause, J., Dasgupta, A., Swartz, J., Aphinyanaphongs, Y., & Bertini, E. (2017). A workflow for visual diagnostics of binary classifiers using instance-level explanations. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 162–172. IEEE.

Lage, I., & Doshi-Velez, F. (2020). Learning interpretable concept-based models with human feedback. *arXiv preprint arXiv:2012.02898*.

Langlotz, C. P., Allen, B., Erickson, B. J., Kalpathy-Cramer, J., Bigelow, K., Cook, T. S., Flanders, A. E., Lungren, M. P., Mendelson, D. S., Rudie, J. D., et al. (2019). A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 nih/rsna/acr/the academy workshop. *Radiology*, *291*(3), 781.

Lee, H., Yune, S., Mansouri, M., Kim, M., Tajmir, S. H., Guerrier, C. E., Ebert, S. A., Pomerantz, S. R., Romero, J. M., Kamalian, S., et al. (2019). An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nature biomedical engineering*, *3*(3), 173–182.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.. *Queue*, *16*(3), 31–57.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, *2*(1), 56–67.

Mahendran, A., & Vedaldi, A. (2015). Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5188–5196.

Martens, D., & Provost, F. (2014). Explaining data-driven document classifications. *MIS quarterly*, *38*(1), 73–100.

Martinez, N., Bertran, M., & Sapiro, G. (2020). Minimax pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*, pp. 6755–6764. PMLR.

Michie, D., Spiegelhalter, D. J., Taylor, C., et al. (1994). Machine learning. *Neural and Statistical Classification*, *13*.

Mincu, D., Loreaux, E., Hou, S., Baur, S., Protsyuk, I., Seneviratne, M., Mottram, A., Tomasev, N., Karthikesalingam, A., & Schrouff, J. (2021). Concept-based model explanations for electronic health records. In *Proceedings of the Conference on Health, Inference, and Learning*, pp. 36–46.

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *11*(3-4), 1–45.

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*, *375*(13), 1216.

Panigutti, C., Guidotti, R., Monreale, A., & Pedreschi, D. (2019). Explaining multi-label black-box classifiers for health applications. In *International Workshop on Health Intelligence*, pp. 97–110. Springer.

Panigutti, C., Perotti, A., & Pedreschi, D. (2020). Doctor xai: an ontology-based approach to black-box sequential data classification explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 629–639.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830.

Plumb, G., Molitor, D., & Talwalkar, A. S. (2018). Model agnostic supervised local explanations. In *Advances in Neural Information Processing Systems*, pp. 2515–2524.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2018). Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). Why should i trust you?: Explaining the predictions of any classifier. In *KDD*.

Saggi, M. K., & Jain, S. (2018). A survey towards an integration of big data analytics to big insights for value-creation. *Information Processing & Management*, *54*(5), 758–790.

Sakr, S., Elshawi, R., Ahmed, A. M., Qureshi, W. T., Brawner, C. A., Keteyian, S. J., Blaha, M. J., & Al-Mallah, M. H. (2017). Comparison of machine learning techniques to predict all-cause mortality using fitness data: the henry ford exercise testing (fit) project. *BMC medical informatics and decision making*, *17*(1), 174.

Sakr, S., & Zomaya, A. Y. (Eds.). (2019). *Encyclopedia of Big Data Technologies*. Springer.

Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., & Batra, D. (2016). Grad-cam: Why did you say that?. *arXiv preprint arXiv:1611.07450*.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*.

Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*, pp. 359–380. PMLR.

Wang, J., Pan, M., He, T., Huang, X., Wang, X., & Tu, X. (2020). A pseudo-relevance feedback framework combining relevance matching and semantic matching for information retrieval. *Information Processing & Management*, *57*(6), 102342.

White, A., & Garcez, A. d. (2019). Measurable counterfactual local explanations for any classifier. *arXiv preprint arXiv:1908.03020*.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer.

Zhang, H., Lu, A. X., Abdalla, M., McDermott, M., & Ghassemi, M. (2020). Hurtful words: quantifying biases in clinical contextual word embeddings. In *proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 110–120.

Zhang, Z., Xie, Y., Xing, F., McGough, M., & Yang, L. (2017). Mdnet: A semantically and visually interpretable medical image diagnosis network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6428–6436.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929.

Zhou, B., Sun, Y., Bau, D., & Torralba, A. (2018). Interpretable basis decomposition for visual explanation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 119–134.

Zucco, C., Liang, H., Di Fatta, G., & Cannataro, M. (2018). Explainable sentiment analysis with applications in medicine. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1740–1747. IEEE.