Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)
Journal Track

Interpretable Local Concept-based Explanation with Human Feedback to Predict All-cause Mortality (Extended Abstract)*

Radwa El Shawi[1] , **Mouaz Al-Mallah**[2]
[1]Institute of Computer Science, Tartu University, Estonia
[2]Houston Methodist DeBakey Heart & Vascular Center
Houston, TX, USA
radwa.elshawi@ut.ee, mal-mallah@houstonmethodist.org

Abstract

Machine learning models are incorporated in different fields and disciplines, some of which require high accountability and transparency, for example, the healthcare sector. 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.

1 Introduction

Machine learning (ML) models have proven to be successful in many application domains, including financial systems, healthcare, agriculture, and criminal justice, especially with the advent of deep learning [Wang *et al.*, 2020; Komisarenko *et al.*, 2022; Shawi *et al.*, 2022]. 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 [Gilpin *et al.*, 2018], 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 mainly focuses on user data protection and privacy; it is expected to cover more algorithmic transparency and explanations requirements from artificial intelligence systems [Goodman and Flaxman, 2017].

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 [Kim *et al.*, 2018].

We summarize our contributions as follows:

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

2 Framework for Local Model Agnostic Concept-based Interpretability

The process of explaining individual predictions is illustrated in Figure 1.

2.1 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

*This paper was originally published in Journal of Artificial Intelligence Research [Shawi and Al-Mallah, 2022]

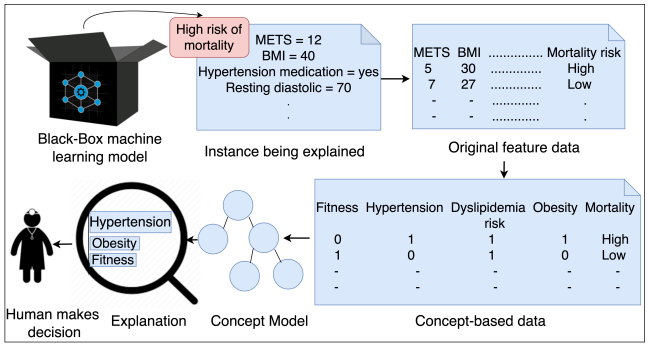

Figure 1: Explaining individual predictions. The patient being explained is represented in low-level features, including vital signs, diagnosis and clinical laboratory measurements. A black-box model predicts this patient as a high risk of mortality. On similar instances to the one being explained, CLEF 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 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.

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:

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

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.

2.2 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 $\pi_{x'}$.

2.3 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. 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. 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 and 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. 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

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 if
    $l_{j^*} = l_{j^*} \cup \{i^*\}$
    $u_{j^*} = u_{j^*} \backslash \{i^*\}$
    $k \leftarrow k + 1$
  end while
  $J^* \leftarrow J^* + 1$
end while

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. 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:

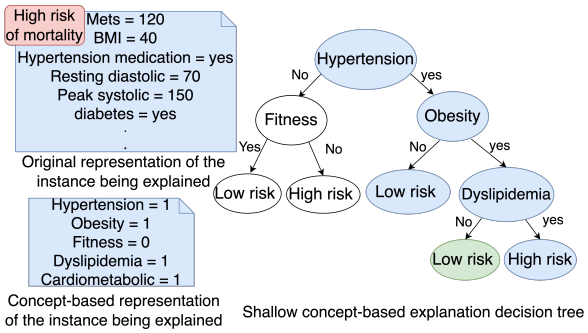

Figure 2: Shallow concept-based explanation decision tree of depth 4 explaining the prediction of a patient with a high risk of mortality

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

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

2.4 Constructing Local Explanation

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'. For the sake of interpretability, we used a shallow decision tree of depth 4 to be comprehensible by humans, leaving the exploration of dynamic depth to future work. Figure 2 shows a decision tree explanation of a patient of a 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'. 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. We do the following to generate a counterfactual explanation from the logistic regression model. Let x'' be the concept representation of the instance being explained x' in terms of high-level concepts. 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'.

3 Results

3.1 Concepts Definition

The dataset used in this work was collected from patients who underwent treadmill stress testing at Henry Ford Affiliated Hospitals, FIT Project [Al-Mallah and others, 2014]. The dataset 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 assumed that the user would accept the proposal of features belonging to these concepts. We relied on clinicians to define a set of handcrafted concepts and associate the ground truth features to each concept. 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.

3.2 Baselines

To explain individual prediction, we compare CLEF to two baselines. The first is an interactive concept-based baseline, and the other 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, equivalent to user feedback on feature-concept association in our approach. Specifically, the interactive baseline, denoted AL, fits five concept-classifier models (regularized logistic regression models), one for each concept. 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 $D_c^+ = \{(x_1, y_c^1), ..., (x_q, y_c^q)|_{y_c=1}\}$ and $D_c^- = \{(x_1, y_c^1), ..., (x_q, y_c^q)|_{y_c=0}\}$, 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. We compare to regularized logistic regression (LR). We train all approaches using the scikit-learn implementations [Pedregosa *et al.*, 2011].

Variant	Downstream accuracy	Concept accuracy
CLIF when f is logistic regression	87% $\pm$0.001	98% $\pm$0.002
CLIF when f is decision tree	88% $\pm$0.001	98% $\pm$0.002
AL	77% $\pm$0.001	85% $\pm$0.003
LR	67% $\pm$0.00	-

Table 1: Downstream and concept accuracies on the testing dataset $\pm$ standard deviation for our proposed technique and baselines.

3.3 Comparison to Baselines

As a black-box model to be explained, we train a random forest model on the training dataset. 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 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 *et al.*, 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 a 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 $98\% \pm 0.002$, which is 13% greater than the AL baseline. This substantial difference suggests that our proposed approach aligns much better with clinician's intuitive representation than the baseline. Our approach with the two variants of f (logistic regression and decision tree) outperforms 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 an 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 collinearity.

References

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

[Gilpin *et al.*, 2018] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[Goodman and Flaxman, 2017] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57, 2017.

[Kim *et al.*, 2018] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[Komisarenko *et al.*, 2022] Viacheslav Komisarenko, Kaupo Voormansik, Radwa Elshawi, and Sherif Sakr. Exploiting time series of sentinel-1 and sentinel-2 to detect grassland mowing events using deep learning with reject region. *Scientific Reports*, 12(1):983, 2022.

[Lage and Doshi-Velez, 2020] Isaac Lage and Finale Doshi-Velez. Learning interpretable concept-based models with human feedback. *arXiv preprint arXiv:2012.02898*, 2020.

[Pedregosa *et al.*, 2011] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[Sakr *et al.*, 2017] Sherif Sakr, Radwa Elshawi, Amjad M Ahmed, Waqas T Qureshi, Clinton A Brawner, Steven J Keteyian, Michael J Blaha, and Mouaz H Al-Mallah. 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, 2017.

[Shawi and Al-Mallah, 2022] Radwa EL Shawi and Mouaz H Al-Mallah. Interpretable local concept-based explanation with human feedback to predict all-cause mortality. *Journal of Artificial Intelligence Research*, 75:833–855, 2022.

[Shawi *et al.*, 2022] Radwa El Shawi, Khatia Kilanava, and Sherif Sakr. An interpretable semi-supervised framework for patch-based classification of breast cancer. *Scientific Reports*, 12(1):16734, 2022.

[Wang *et al.*, 2020] Junmei Wang, Min Pan, Tingting He, Xiang Huang, Xueyan Wang, and Xinhui Tu. A pseudo-relevance feedback framework combining relevance matching and semantic matching for information retrieval. *Information Processing & Management*, 57(6):102342, 2020.