# Concept-Level Model Interpretation From the Causal Aspect

Liuyi Yao, Yaliang Li, Sheng Li, *Senior Member, IEEE*, Jinduo Liu, Mengdi Huai, Aidong Zhang, *Fellow, IEEE*, and Jing Gao

**Abstract**—With the increasing growth of data and the ability of learning with them, machine learning models are adopted in various domains. However, few of machine learning models are able to reason their prediction, which limits their further applications in real-world tasks. With the potential to address this dilemma, model interpretation has become an important research topic because of the ability to provide the underlying reasons for model predictions at the feature level or concept level. Model interpretation at the concept level focuses on exploring the roles of concepts in model prediction, which enables more compact and understandable interpretations. Concept-level model interpretation requires the identification of the concepts that contribute to model prediction and the exploration of the rules underneath these concepts. To achieve the two objectives, we propose a Concept-level Model Interpretation framework (CMIC) from the perspective of causality. CMIC can automatically detect concepts in data and discover the causal relation between the detected concepts and the model's predicted labels. Furthermore, CMIC ranks the contributions of concepts by their causal effect on the model prediction, reflecting the detected concepts' importance. We evaluate the proposed CMIC framework on both synthetic and real-world datasets to demonstrate the quality of the provided interpretation.

**Index Terms**—Model interpretation, causal discovery

## 1 Introduction

Recently, with the generation of an enormous amount of data, machine learning models are popular in various domains due to their ability to learn from data. In the real-world applications, knowing the reasons behind the machine model prediction is critical for people to decide whether to trust the model. Especially in the high-risk domains, such as medicine and finance, providing the reasons for prediction is highly desired for safe and broad applications of machine learning models. Owing to the ability to reveal the inner mechanism of machine learning models [2], [7], [19], model interpretation has become a trending topic in recent years. Moreover, model interpretation is important for a model to be accepted by real-world applications. In turn, it further facilitates the model design and debugging when diving into the reasons behind model predictions to inspect the model.

Most existing works on model interpretation focus on single feature level interpretation. Existing methods assign each underlying feature an importance score, indicating the key features for model prediction [7], [36]. The representative scoring methods include the gradient-based scores [36], [38], Shapley value-based scores [2], [7], [27], [29], [36], [41], and perturbation based scores [11]. However, interpreting models at the single feature level suffers from some limitations. First, a single feature may lack semantic meanings. For example, a single pixel in images, a single word in documents, or a single value in the gene expression data may not correspond to meaningful semantics. Second, in high dimensional data, the feature importance vector would be large, which makes it difficult for a human to understand. Inspecting the importance vector of all those features is time-consuming, and it could be challenging to infer a proper interpretation. Third, when handling high dimensional data, the features may be noisy or contain redundant information, which makes the single feature level interpretation vulnerable.

As complementary to the single feature interpretation, the interpretation at the concept level can overcome the aforementioned limitations. Concept, an intermediate-level summarization of data, is more concretized than a single feature, making it more readable for humans. For example, in medical datasets, a combination of features, such as patients' residence, yearly income, occupation, and education level, compose a socio-economic status concept, which is easy to interpret. Moreover, as a summarization of the original data, the concept can filter out redundant information and be less sensitive to noise. To conduct model interpretation at the concept level, the following two questions need to be answered: *What concepts contribute to the model prediction? What are their roles in the model prediction?*

A few concept-level interpretation methods have been proposed in the literature to answer the above two questions.


- *Liuyi Yao and Yaliang Li are with Alibaba Group, Hangzhou 311121, China. E-mail: {yly287738, yaliang.li}@alibaba-inc.com.*
- *Sheng Li and Aidong Zhang are with the University of Virginia, Charlottesville, VA 22904 USA. E-mail: {shengli, aidong}@virginia.edu.*
- *Jinduo Liu is with the Beijing University of Technology, Beijing 100021, China. E-mail: jinduo@bjut.edu.cn.*
- *Mengdi Huai is with Iowa State University, Ames, IA 50011 USA. E-mail: mdhuai@iastate.edu.*
- *Jing Gao is with Purdue Univeristy, West Lafayette, IN 47907 USA. E-mail: jinggao@purdue.edu.*


*Manuscript received 19 December 2021; revised 15 July 2022; accepted 9 September 2022. Date of publication 27 September 2022; date of current version 8 August 2023.*
*This work was supported in part by the National Science Foundation under Grants NSF IIS-2141037 and IIS-2226108.*
*(Corresponding author: Liuyi Yao.)*
*Recommended for acceptance by N. Chawla.*
*Digital Object Identifier no. 10.1109/TKDE.2022.3209997*

The quantitative testing with concept activation vectors (TCAV) [24], and automatic concept-based explanations (ACE) [15] learn the representations of the pre-defined concepts in the original data space, and adopt gradient-based score methods to measure the importance of such concepts. The causal concept effect (CaCE) method [17] adopts an intervention-based strategy, which modifies the original data by forcing them to contain or not contain one specific concept, and the importance of such concept is measured as the difference in label predictions after the intervention. Although promising, these existing methods still have some limitations. Both TCAV and CaCE require prior knowledge about the concept. Besides, TCAV and ACE generate model-specific interpretations, which require access to the structures and parameters of the model. Furthermore, CaCE relies on pre-defined causal relations between the concepts and the model predictions to calculate the importance score of a concept, which might obtain unreliable results when the assumption about the causal relation is not satisfied.

In light of the above challenges, we propose a **C**oncept-level **M**odel **I**nterpretation framework from the **C**ausal aspect, abbreviated as *CMIC*. CMIC automatically extracts potential concepts in the available data and meanwhile provides readable descriptions of the extracted concepts. To explore what concepts contribute to the model prediction and analyze their importance, the relationship between the extracted concept and the model's predicted labels is studied from a causal aspect. A concept contributes to the model prediction if it is a cause of the predicted label. The causal effect of a concept on the predicted label is viewed as the importance score of this concept. Our CMIC framework consists of three components. In the first component, concepts are extracted in an unsupervised way, along with the generation of understandable descriptions for each extracted concept. In the second component, a causal graph for the extracted concepts and the predicted labels is constructed to identify concepts that contribute to the model prediction. The third component is the causal effect analysis. The importance score of the identified concept is calculated, which is regarded as the causal effect of such a concept on the model predicted labels. Experiments on both synthetic and real-world datasets show that the proposed CMIC framework can generate meaningful concept-level model interpretations, which provides a lens to explain the performance difference of different classifiers.

The rest of this paper is organized as follows. In Section 2, we discuss of the related work on model interpretation. Section 3 presents an overview of the propose CMIC frameworks. In Section 4, the details of the proposed CMIC framework are presented. Section 5 introduces the experiments on both synthetic and real-world datasets. Finally, Section 6 concludes this paper and points out the future directions.

## 2 Related Work

We summarize the related work into four categories: (1) Local interpretation; (2) Global interpretation; (3) Counterfactual interpretation; (4) Concept-based interpretation.

*Local Interpretation.* In recent years, various local interpretation methods have been proposed to provide explanations for classification models through scoring the importance of each input feature for a given instance[2], [7], [9], [18], [26], [27], [29], [34], [38], [41]. The authors in [34] propose LIME (Local Interpretable Model-Agnostic), an interpretation method that explains the predictions of any classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction. Besides, a family of quantitative input influence measures that capture the degree of influence of inputs on outputs of systems is introduced [9]. DeepLIFT ((Deep Learning Important FeaTures) [38] is proposed for decomposing the output prediction of a neural network on a specific input by backpropagating the contributions of all neurons in the network to every feature of the input. In addition, the Shapley-value-based methods have been proposed to provide local interpretations, by assigning each feature an importance value for a particular prediction [2], [7], [27], [29], [36], [41].

*Global Interpretation.* In contrast with local interpretation methods that only capture the local behavior of the model on a local region of the input space, global explanation methods [19], [22], [28], [33], [34], [42], [45] aim to explain the overall decision-making process of a model. Some methods in this category provide global explanations via the surrogate models [28], [33], [45]. For example, the authors in [33] propose to learn if-then rules to globally explain the behavior of black-box models that have been used to solve classification problems. There are also some other global interpretation methods [19], [22], [34], [42] that can provide explanations for different populations. In [42], model distillation is leveraged to learn global additive explanations that describe the relationship between input features and model predictions. In [22], the authors provide a global attribution method by grouping local features with similar importance scores. In [34], the global interpretation is constructed by aggregating the weights of linear models. In [19], the authors use an enhanced mixture model to approximate the target model, and then extracts the global interpretations from the derived enhanced mixture model.

*Counterfactual Interpretation.* Counterfactual has been extensively discussed in the causal inference literature [32]. Recently, some counterfactual explanation methods [1], [40] have been proposed to explain predictions of individual instances. The authors in [40] show example explanations, discuss their strengths and weaknesses, illustrate how they can be used to debug the underlying model, inspects its fairness, and also unveils security and privacy challenges that they pose. Moreover, CoCoX (shorted for Conceptual and Counterfactual Explanations), introduced in [1], can explain decisions made by a convolutional neural network (CNN) using fault-lines. Specifically, given an input image for which a CNN model predicts a class, the proposed fault-line based explanation can identify the minimal semantic-level features (referred to as explainable concepts).

*Concept-Based Interpretation.* By far, there are some concept-based interpretation methods have been proposed [13], [14], [15], [20], [30], [35], [46]. The authors in [24] lay out the general principles and desiderata for the concept-based explanation, and then proposed TCAV method, which tests the concept activation vectors to reflect the concept importance. Further, based on TCAV, a systemic framework ACE [15], is developed to identify higher-level concepts that are meaningful

TABLE 1
Comparison Between CMIC and Existing Concept-Level Interpretation Methods

| Method | Automatic Concept Detection | Explore Causal Relationship | Causal Graph Learning |
|---|---|---|---|
| TCAV | ✗ | ✗ | ✗ |
| ACE | ✓ | ✗ | ✗ |
| CACE | ✗ | ✓ | ✗ |
| AutoRMI | ✓ | ✗ | ✗ |
| CMIC | ✓ | ✓ | ✓ |

to humans. In [46], concept-based explainability for DNNs (Deep Neural Networks) is studied in a systematic framework, and proposes a concept discovery method that considers two additional constraints to encourage the interpretability of the discovered concepts. Further, in [13], the authors improve the interpretability of a similarity learning system, and designs a deep interpretable architecture for similarity learning built upon hierarchical concepts. CaCE [17] examines the importance of the concept by comparing the prediction difference on the data with or without such a concept. In [14], the authors provide node-level concept-based reasoning for graph neural network (GNN) models by introducing Concept Bottleneck Graph Neural Networks (CBGNNs). In [20], the authors propose a automatic and robust model interpretation method (AutoRMI), which automatically generates the prototype-based concept explanations with certified robustness guarantees. In [35], the authors propose a framework that can add to any backbone neural network to jointly learning to predict and generate the ante-hoc explanations via concepts.

Compared with the existing concept-level interpretation methods, the proposed CMIC framework works for blackbox models, which is a significant difference to model-specific interpretations [15], [24], [46]. Besides, different from CaCE [17] and TCAV [24] that require concept specification, CMIC is able to detect the concepts and express readable concept meanings automatically. Another significant difference between the proposed CMIC and CaCE is that CaCE predefines the causal graph, which may not always be faithful to the actual causal graph, and CMIC avoids this drawback by discovering causal relations from data. Overall, We compare our work with existing works including TCAV, ACE, CACE, and AutoRMI in terms of the following three aspects: (1) whether it can automatically extract the concept, (2) whether it explore the causal relationship between the extracted concepts and the prediction, (3) whether it learns the causal graph between the extracted concepts and the prediction. The comparison between CMIC and existing concept-level interpretation methods is summarized in Table 1.

# 3 Overview

## 3.1 Problem Definition

The studied problem is to interpret a target classification model, denoted as $f$, at the concept level. The input of the proposed CMIC framework includes a sample set denoted as $\mathbf{X}$, and the output labels of model $f$, denoted as $L_f$, where $\mathbf{X} \in \mathcal{R}^{n\times d}$, $n$ is the number of samples in $\mathbf{X}$, $d$ is the number of features, and $L_f = f(\mathbf{X}) \in \mathcal{R}^n$. For presentation clarity, we name $L_f$ as the $f$**-label**. The output of our framework is the concept-level interpretation for the target model $f$, including a set of $N_c$ concepts $\{A_i\}_{i=1}^{N_c}$ with each concept associated with human-friendly concept meanings, a causal graph $\mathcal{G}$ which shows the causal relation between the extracted concepts and $L_f$, and the importance scores of concepts that are relevant to the model $f$.

## 3.2 Proposed Framework

Fig. 1 shows the framework of our proposed CMIC interpretation method, which contains three steps. The first step is concept extraction, in which the potential concepts in the feature data $\mathbf{X}$ are extracted in an unsupervised way. In order to identify model-$f$ related concepts, our second step, denoted as concept-label causal relation discovery, aims to explore the relationship between the extracted concepts and the $f$-label. Naturally, those model-$f$ related concepts identified in the second step are fed into the last step called concept effect analysis, whose objective is to understand the significance of the identified concepts in model-$f$'s label prediction. The following section introduces the three steps at length.

# 4 Methodology

## 4.1 Concept Mining

How to quantitatively define and extract concepts from data, such as images and text, has been an active research topic for decades [23]. In this work, we focus on extracting concepts from structured data. Specifically, *concept* is defined as some common characteristics shared by a subset of samples in the dataset. Based on the definition of the concept, the samples containing the same concept can be viewed as a cluster. Therefore, in the first stage of concept extraction, CMIC explores the discriminative clusters in the dataset as much as possible. Next, the concept meaning contained in each cluster will be extracted to illustrate the concept quantitatively.

### 4.1.1 Discriminative Clustering for Concept Detection

To fulfill the requirement of exploring as many discriminative clusters as possible, we adopt the discriminative clustering [39] method described as follows. The entire dataset $\mathbf{X}$ is separated into two sets, i.e., a "discovery dataset" $\mathcal{D}$ and a "natural dataset" $\mathcal{N}$. The discovery dataset aims to discover all potential clusters, and the natural dataset is an auxiliary source to ensure the discovered clusters are discriminative. In detail, initial clusters in the discovery dataset are estimated by the $K$-means clustering. For each cluster whose size is larger than a pre-defined parameter $k$, a binary SVM (Support Vector Machine) classifier [43] is trained by considering this cluster as the positive class and the natural dataset as the negative class. After training the SVM classifier on the combined dataset, the top $m$ samples with the highest SVM scores in the discovery dataset are used to form a new cluster associated with that SVM classifier. The above two procedures, SVM classifier training and label assignment on the discovery dataset, are repeated until convergence.

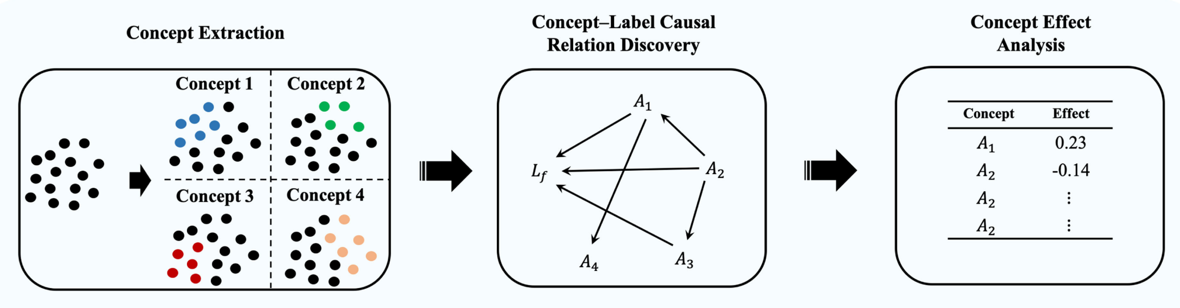



Fig. 1. The framework of CMIC. The proposed CMIC framework contains three steps. The first step is concept extraction, which automatically discovers the potential concepts and transforms the original data into concept-level data. The second step is causal structure learning, which explores the causal relationship between the extracted concepts ($A_1, A_2, \ldots, A_4$) and the $f$-label $L_f$. The last step is the Concept Effect Analysis, which measures the importance of the concept from the causal view.

The motivation of using the SVM classifier on the combined dataset is to ensure that the detected cluster in the discovery dataset is discriminative to the whole dataset. Therefore, clusters detected by the discriminative clustering perfectly fit our requirements.

### 4.1.2 Concept-Level Data Transformation

After discriminative clustering, we can obtain totally $N_c$ SVM classifiers, and each classifier is a detector for one specific concept. One sample $x \in \mathcal{R}^d$ contains the $i$-th concept if $S_i(x) = 1$, where $S_i(\cdot)$ denotes the $i$-th SVM classifier.

By utilizing the obtained SVM classifiers, the original data can be transformed into concept-level data. Let $\mathbf{A} \in \mathcal{R}^{n \times N_c}$ denote the transformed concept data, and $A_{i,j} = S_j(x_i)$, where $S_j$ is the $j$-th SVM classifier, and $x_i$ is the $i$-th sample of $\mathbf{X}$. In other words, $A_{i,j} \in \{0,1\}$ indicates whether the $i$-th sample $x_i$ contains the $j$-th concept or not.

After obtaining the concept detectors and transforming the data into a concept space, the next stage is to quantitatively explore the semantic meaning of each extracted concept.

### 4.1.3 Concept Meaning Extraction

Our concept meaning extraction method is motivated by the fact that one concept can be expressed by the combination of its proxy variables. For example, the concept "*good socio-economic status*" can be expressed as the features "*yearly income*" $> 200K$ and *"residence" in wealthy neighborhoods (e.g., Los Altos Hills in California)*. We assume that the meaning of each concept is a subset of features with certain value ranges. Based on this assumption, we propose a two-step procedure for concept meaning extraction: (1) Select a subset of features; (2) Determine the value range of each selected feature.

*Step 1: Feature Selection.* A feature selected to describe the concept meaning should satisfy the following criteria. (1) Within the cluster, the values of feature should be homogeneous. (2) Across clusters with different concepts, the values should be heterogeneous. To fulfill the requirements, the following is proposed for feature selection.

Let $S_i$ denote the SVM classifier of the $i$-th concept. Let $X^{(i)}$ denote the positive sample set with respect to the $i$-th concept, i.e., $X^{(i)} = [(x_1^{(i)})^T, (x_2^{(i)})^T, \ldots, (x_{N_i}^{(i)})^T]^T$, where $X^{(i)} \in \mathbb{R}^{N_i \times d}$, $x_j^{(i)} \in \mathbf{X}$ satisfies $S_i(x_j^{(i)}) > 0.5$, for $j = 1, 2, \ldots, N_i$. $S_i(x_j^{(i)})$ is the output of the $i$-th SVM classifier, which is the probability of containing the $i$-th concept. $N_i$ is the number of samples in $X^{(i)}$ and $d$ is the total number of features in the original dataset $\mathbf{X}$. Implementing the second criterion involves other concepts' clusters. To this end, we construct a negative sample set for the $i$-th concept, denoted as $X^{(-i)}$, and

$$X^{(-i)} = \left[ \left(x_1^{(-i)}\right)^T, \left(x_2^{(-i)}\right)^T, \ldots, \left(x_{N_i}^{(-i)}\right)^T \right]^T,$$

where $X^{(-i)} \in \mathbb{R}^{N_i \times d}$, $x_1^{(-i)}, x_2^{(-i)}, \ldots, x_{N_i}^{(-i)}$ are the top $N_i$ samples with $S_i(x_j^{(-i)}) < 0.5$, $j = 1, 2, \ldots, N_i$.

Based on our design criteria, we propose the following objective function to select features that form a concept.

$$\begin{aligned} \max_{p_k^{(i)}} & \sum_{i=1}^{N_c} \sum_{k=1}^{d} p_k^{(i)} \left( d_{\text{cross}}(X_k^{(i)}, X_k^{(-i)}) + d_{\text{within}}(X_k^{(i)}) \right) \\ & - \lambda \sum_{k=1}^{d} 1\!\!1_{\{p_k^{(i)} > 0.5\}} \\ & \text{s.t. } 0 \le p_k^{(i)} \le 1, \end{aligned} \quad (1)$$

where $p_k^{(i)}$ denotes the probability that the $k$-th feature is selected to form the concept in the $i$-th cluster. $X_k^{(i)}$ denotes the vector of the $k$-th feature in $X^{(i)}$, $d_{\text{cross}}(\cdot,\cdot)$ denotes the cross concept distance, and $d_{\text{within}}$ measures the homogeneity of the feature within the concept. $1\!\!1_{\{\cdot\}}$ is an indicator function, and $\lambda$ is a hyperparameter. By maximizing the first term, features with high heterogeneity across the clusters and high homogeneity within the cluster will be selected. The second term is a regularization term restricting the number of selected features.

To make the Eqn. (1) differentiable, the Wasserstein distance [8], [44] is adopted to measure the heterogeneity across the concept, and the variance is used to measure the homogeneity within the concept. Besides, a differentiable approximate function to the regularization term is also adopted. Overall, the transformed objective is shown as follows:

$$\begin{aligned} \max_{p_k^{(i)}} & \sum_{k=1}^{K} p_k^{(i)} \left( \text{WASS}(X_k^{(i)}, X_k^{(-i)}) - \text{var}(X_k^{(i)}) \right) \\ & - \lambda ||f_{ap}(P)||_F \\ & \text{s.t. } 0 \le p_k^{(i)} \le 1, \end{aligned} \quad (2)$$

where $\text{WASS}(\cdot,\cdot)$ is the Wasserstein distance; $\text{var}(\cdot)$ is the variance; $f_{ap}(P)$ is the auxiliary approximation function with $f_{ap}(x) = \frac{1}{(1+\exp(-v(x+\frac{1}{\sqrt{v}})))(1+\exp(-v(1-x+\frac{1}{\sqrt{v}})))}$, where $v$ is a scalar value determining the approximate level; $P$ is the feature selection probability matrix, $P_{ij} = p_j^{(i)}$; $||\cdot||_F$ is the Frobenius norm.

By solving the transformed objective function Eqn. (2), we can obtain the feature selection probability matrix $P$. If $p_k^{(i)} > 0.5$, the $k$-th feature is selected as the forming feature of the $i$-th concept.

*Step 2: Value-Range Determination.* This step determines the value ranges associated with the selected features to generate human-friendly concept meaning. Suppose the $k$-th feature is selected as the component of the $i$-th concept, and its positive sample set is $X_k^{(i)}$. If the selected feature is categorical, the item that appears most frequently in $X_k^{(i)}$ is the value of this feature in the $i$-th concept. The quantile statistics are adopted if the selected feature is ordinal (either continuous or discrete). As the positive samples might be noisy, using quantile statistics can avoid this issue to some extent. Then, the value range is defined as an interval: $[25\% \text{ percentile of } X_k^{(i)}, 75\% \text{ percentile of } X_k^{(i)}]$.

### 4.2 Causal Structure Learning

Concepts are extracted in the previous step; however, not all concepts contribute to the label prediction. This subsection aims to explore the causal relationships between the concepts and the model predicted labels, and select the important concepts for further analysis.

The important concepts contribute to the label prediction, leading to a causal relationship between the important concepts and the predicted label. Therefore, in this section, we explore the causal structure between the concepts and $f$-label. One effective approach for exploring the causal structure between variables is the causal discovery model. In general, causal discovery estimates a directed acyclic graph (DAG) from data, which reflects the causal relationships between variables. Let $\mathcal{G}$ denote a directed acyclic graph. A causal graph can be expressed as $\mathcal{G} = < \mathbf{V}, \mathbf{E} >$ , where $\mathbf{V}$ is a set of nodes representing the extracted concepts and $f$-label, i.e., $\mathbf{V} = \{A_1, A_2, \cdots, A_{N_C}, L_f\}$, $A_i$ is the $i$-th concept, $L_f$ is the $f$-label, and $\mathbf{E}$ is a set of arcs with each arc $V_i \rightarrow V_j$ ($A_i \rightarrow A_j$ or $A_i \rightarrow L_f$) describing a causal relation between two nodes. For notation clarity, we use $V_i$ to denote the $i$-th element in $\mathbf{V}$. In summary, the inputs of our causal discovery model are the transformed concept-level data $\mathbf{A}$ and $f$-labels, denoted as $\mathcal{D}_c$, where $\mathcal{D}_c \in \mathcal{R}^{n\times(N_c+1)}$ is a concatenation of $\mathbf{A}$ and $L_f$, and the output is the causal graph $\mathcal{G}$ indicating the causal relationships among $\{A_1, A_2, \ldots, A_{N_c}, L_f\}$.

Causal structure learning aims at learning a causal graph and ensuring all the directions of the causal graph are determined. In other words, the learned graph is a causal graph instead of a Bayesian network or Markov equivalent classes. Therefore, we adopt the Structural equational likelihood framework (SELF) [5], which dissolves the ambiguity from the Markov equivalent classes, and provides a unified and theoretically robust methodology for causal structure exploration. SELF focuses on the noise estimation, by maximizing the global likelihood of the entire Bayesian network while preserving local statistical independence between noise and cause variables.

In detail, for a node $V_i \in \mathbf{V}$, it can be presented by the causal mechanism: $V_i = F_i(\Pi(V_i)) + e_i$, where $F_i$ is the causal function of $V_i$, $\Pi(V_i)$ is the parent nodes of $V_i$, and $e_i$ is the randomized noise which is independent of $\Pi(V_i)$ ($e_i \perp\!\!\!\perp \Pi(V_i)$). Then given the data $\mathcal{D}_c$, we can construct a causal graph $\mathcal{G}$ and corresponding structural equations $F_i$ for all variables in $A_i$ and $L_f$ by maximizing the score function, which is defined as:

$$S(\mathcal{G}, \mathcal{D}_c) = \sum_{i=1}^{N_c+1} \log\left(P(e_i = V_i - F_i(\Pi(V_i)))\right) - \frac{d_p}{2}\log n, \tag{3}$$

where $\frac{d_p \log(n)}{2}$ is a penalty, $d_p$ is the number of total coefficients used in $\{F_i\}_{i=1}^{N_C+1}$. In particular, $L_f \notin \Pi(A_i)$, which means that the $f$-label can not be the cause of concepts.

After the causal discovery on the transformed dataset, the learned causal graph directly reveals the causal relationships between the concepts and the $f$-labels. The concepts, which have the causal path to the $f$-label $L_f$, are selected as important concepts for the effect analysis in the next subsection.

### 4.3 Concept Effect Analysis

Analyzing the effect of concepts helps understand the different roles that the concepts play in the target model's label prediction. With the causal graph available, Pearl's graphical causal model (GCM) [32] is adopted to measure the causal effect of concepts on the model's predicted labels. The core of GCM is the intervention, which, in our case, aims to study how the predicted label changes when we forcibly restrict all the samples containing or not containing one specific concept. Mathematically, GCM utilizes the *do*-calculus to model the causal effects. In particular, the adjusted model prediction after intervention on the $i$-th concept is denoted as $p(L_f|do(A_i))$, where $do(A_i) = 1$ means forcibly making all samples contain the $i$-th concept $A_i$, and, similarly, $do(A_i) = 0$ indicates forcing all samples not to contain the $i$-th concept. Based on the intervention, the effect of the $i$-th concept is formulated as:

$$E_{A_i} = p(L_f = 1|do(A_i) = 1) - p(L_f = 1|do(A_i) = 0). \tag{4}$$

Binary labels are considered in Eqn. (4). When there are multiple labels, they can be automatically transformed to binary labels by one-hot encoding, and when the label is continuous, the Eqn. (4) can be transformed into the expectation version:

$$E_{A_i} = \mathbb{E}\left[L_f|do(A_i) = 1\right] - \mathbb{E}\left[L_f|do(A_i) = 0\right].$$

After defining the concept effect by the *do*-calculus, the complete identification algorithm (ID-algorithm) [37] is adopted to transforms the above *do*-calculus expression into a regular probability expression. We use a toy example, whose causal graph is shown in Fig. 2, to illustrate how to identify the effect. In Fig. 2, the second concept $A_2$ has some confounded paths to $L_f$, which means $A_2$ and $L_f$ have some common causes, $A_1$ and $A_3$. Therefore, according to the back-door criteria [32], the probability of $L_f$ after intervention on $A_2$ is formulated as:

$$p(L_f|do(A_2)) = \sum_{A_1,A_3} p(L_f|A_1, A_2, A_3)P(A_1)p(A_3). \tag{5}$$

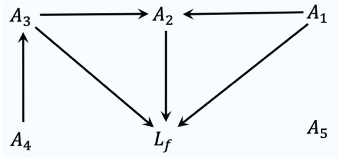



Fig. 2. An example of causal graph.

Compared with $A_2$, there is no confounded path between $L_f$ and $A_1, A_3, A_4$. Thus the probability of $L_f$ after intervention is: $p(L_f|do(A_i)) = p(L_f|A_i)$, where $i = 1, 3, 4$.

Overall, after calculating the effect of each concept on the $f$-label, the model-$f$ related concepts can be ranked based on these effects, which provide concept-level model interpretations from the perspective of causality.

## 5 EXPERIMENTS

In this section, we conduct experiments on synthetic and real-world datasets to validate the following aspects: (1) CMIC can extract high-quality concepts. (2) CMIC can provide concept-level reasoning and interpretations to explain the performance differences of different classifiers.

### 5.1 Experiments on Synthetic Dataset

Since there are no ground truth concepts in the real-world datasets, we experiment on the synthetic dataset, whose data are generated from the pre-defined concepts, to evaluate the concept extraction procedure quantitatively.

#### 5.1.1 Data Generation

The synthetic data generation contains two steps: (1) Concept-level data generation; (2) Feature-level Data Generation.

*Concept-Level Data Generation.* In this step, the concept-level dataset is generated according to the predefined causal graph, shown in Fig. 3. The procedure for generating the four concepts and the label is as follows: $A_1, A_2, A_3, A_4 \sim \text{Bernoulli}(0.5)$, $L_f \sim \text{Bernoulli}(\text{logit}(-A_1+3A_2-2A_3+4A_4))$, where logit denotes the logistic function. After repeating the above procedures 100 times, we obtain the concept-level synthetic data $\mathbf{A} \in \mathcal{R}^{100\times 4}$ and the label vector $L_f \in \mathcal{R}^{100}$.

*Feature-Level Data Generation.* The concept meaning is defined in Table 2, where $d_i$ represents the $i$-th feature, and its value range is specified by its following interval. The concept meanings, along with the concept-level data, determine the value range of each feature in each sample, and thus the feature values can be sampled accordingly.

To better describe the feature-level data generation procedure, the sample, whose concept-level data is [1,0,1,0], is taken as an example. According to Table 2, the value range for each feature is: $\{d_1 : [-1, 1], d_2 : [-5, 3],\ d_3 \notin [10, 14]$ or $d_4 \notin [5, 7], d_5 : [9, 11], d_6 : [10, 14], d_7 \notin [15, 17]\}$. The feature-level data is generated by the following procedures.

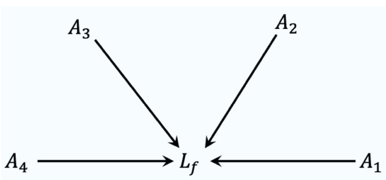



Fig. 3. Causal graph to generate synthetic concept.

TABLE 2
Synthetic Data Generation: Concept Meaning

| Concept | Meaning |
|---|---|
| $A_1$ | $d_1$: $[-1, 1]$, $d_2$: $[-5, 3]$ |
| $A_2$ | $d_3$: $[10, 14]$, $d_4$: $[5, 7]$ |
| $A_3$ | $d_5$: $[9, 11]$, $d_6$: $[10, 14]$ |
| $A_3$ | $d_7$: $[15, 17]$ |

- For the features whose value range does not contain $\notin$ notation, their values are uniformly sampled on the specified interval. In this example, the values of feature $d_1, d_2, d_5, d_6$ are uniformly sampled from intervals $[-1, 1], [-5, 3], [9, 11], [10, 14]$, respectively.
- For the features whose value range contains the $\notin$ notation and the "or" logic, they are first sampled from a predefined range, named as open range. Then, we check whether the generated values satisfy the "or" condition. If not, we repeat the sampling procedure until satisfying. In this example, the values of $d_3$ and $d_4$ are first uniformly sampled from the open range $[-20, 20]$, and then we check whether the sampled values satisfy the "or" condition.
- For the features whose value range only contains $\notin$ condition, their values are uniformly sampled from the open range excluding the interval marked by $\notin$. In this example, the open range is $[-20, 20]$, and the value of $d_7$ is uniformly sampled from the interval $[-20, 15) \cap (17, 20]$.

#### 5.1.2 Experiment Settings

In the following, the baselines and the evaluation metric adopted in the experiment are introduced.

*Baselines.* We compare our proposed concept extraction method with spectral bi-clustering [10], [25], which simultaneously clusters rows and columns of the data matrix. Each cluster of rows and columns determines a sub-matrix of the original data matrix in bi-clustering. Thus, each sub-matrix can be viewed as a concept, and the concept-level data can be acquired accordingly.

*Evaluation Metric.* Since the concept-level transformation is based on the clustering results, evaluation metrics that are commonly used in clustering can be adopted. In this experiment, we adopt the Adjusted Rand Index (ARI) [21] as the evaluation metric, and a higher ARI score indicates a better performance.

#### 5.1.3 Results and Analysis

Fig. 4 shows the results of our proposed CMIC method and the bi-clustering method. The cell of the $i$-th row and the $j$-th column is the ARI score between the $i$-th ground truth concept and the $j$-th extracted concept. In other words, each cell represents the ARI score between the $i$-th column in the ground truth concept-level data $\mathbf{A}$ and the $j$-th column in $\mathbf{A}^*$, where $\mathbf{A}^*$ is the transformed concept-level data either by CMIC or bi-clustering. The left subplot shows the ARI results of our proposed method, and the right one shows the results

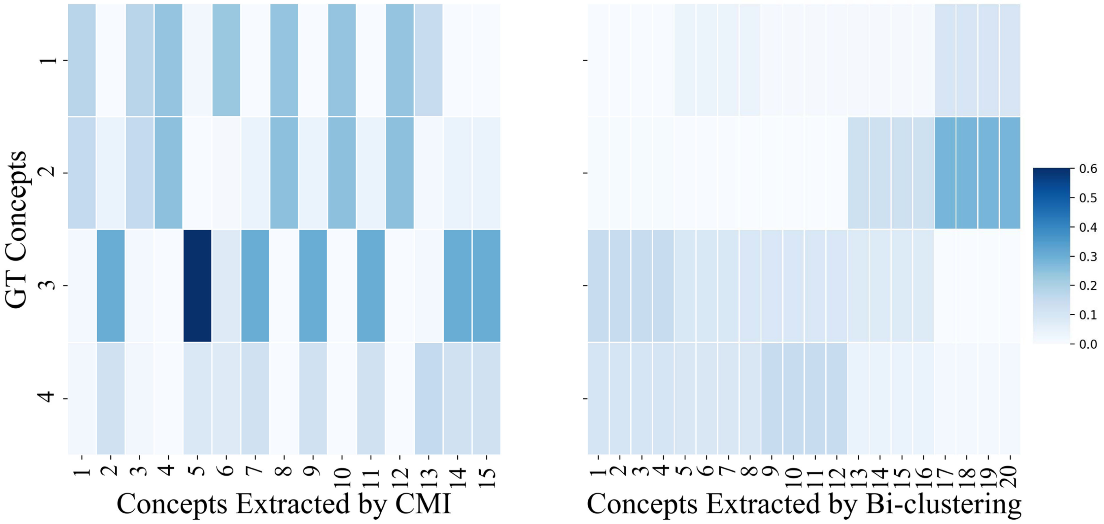



Fig. 4. ARI Score between ground truth concepts and the extracted concepts by different extraction methods.

of the baseline method, where the $y$-axis label "GT Concepts" denotes the ground truth concepts. The hyper-parameter $\lambda$ of CMIC in Eqn. (2) is set as 0.1.

From Fig. 4, it can be observed that the concepts extracted by our proposed method can cover more ground truth concepts than the baseline method of bi-clustering. One possible reason is that, bi-clustering performs hard clustering on all the features, which limits concepts' expressiveness. In contrast, in our proposed method, one feature can be included in multiple concepts, which brings more flexibility to concept extraction.

## 5.2 Experiments on Real-World Datasets

In this section, we experiment on two real-world datasets to qualitatively examine the following: (1) The extracted concepts are meaningful; (2) Explain why different classifiers perform differently at the concept level.

### 5.2.1 Dataset

In this experiment, two publicly available real-world datasets are adopted, including the Bank Marketing dataset and Divorce dataset.

*Bank Marketing*. This dataset is first introduced in [31], which records results of the direct marketing campaigns (phone call) on 4522 clients.[1] In this dataset, there are 17 attributes related to clients' demographic information such as age, job, marital status, phone call duration, previous and current campaign information, and the outcome of the previous campaign. The class label is binary, indicating whether the product (bank term deposit) was subscribed.

*Divorce*. This dataset was collected in a study about divorce [47]. In the study, divorced couples and couples with happy marriages are required to answer 54 questions, scaled from 0 to 5, related to their marriage. The classification label is binary, indicating whether they are divorced or married couples. Overall, there are 170 records available in the dataset.[2]

1. https://archive.ics.uci.edu/ml/datasets/Bank+Marketing
2. https://archive.ics.uci.edu/ml/datasets/Divorce+Predictors+data+set

TABLE 3
F1 Score of Models on Real-World Datasets

| | SVM | DT | RF | NN | Ada |
|---|---|---|---|---|---|
| Bank Marketing | 0.47 | 0.41 | 0.45 | 0.41 | 0.39 |
| Divorce | 0.42 | 0.89 | 0.98 | 0.98 | 0.98 |

### 5.2.2 Experimental Settings

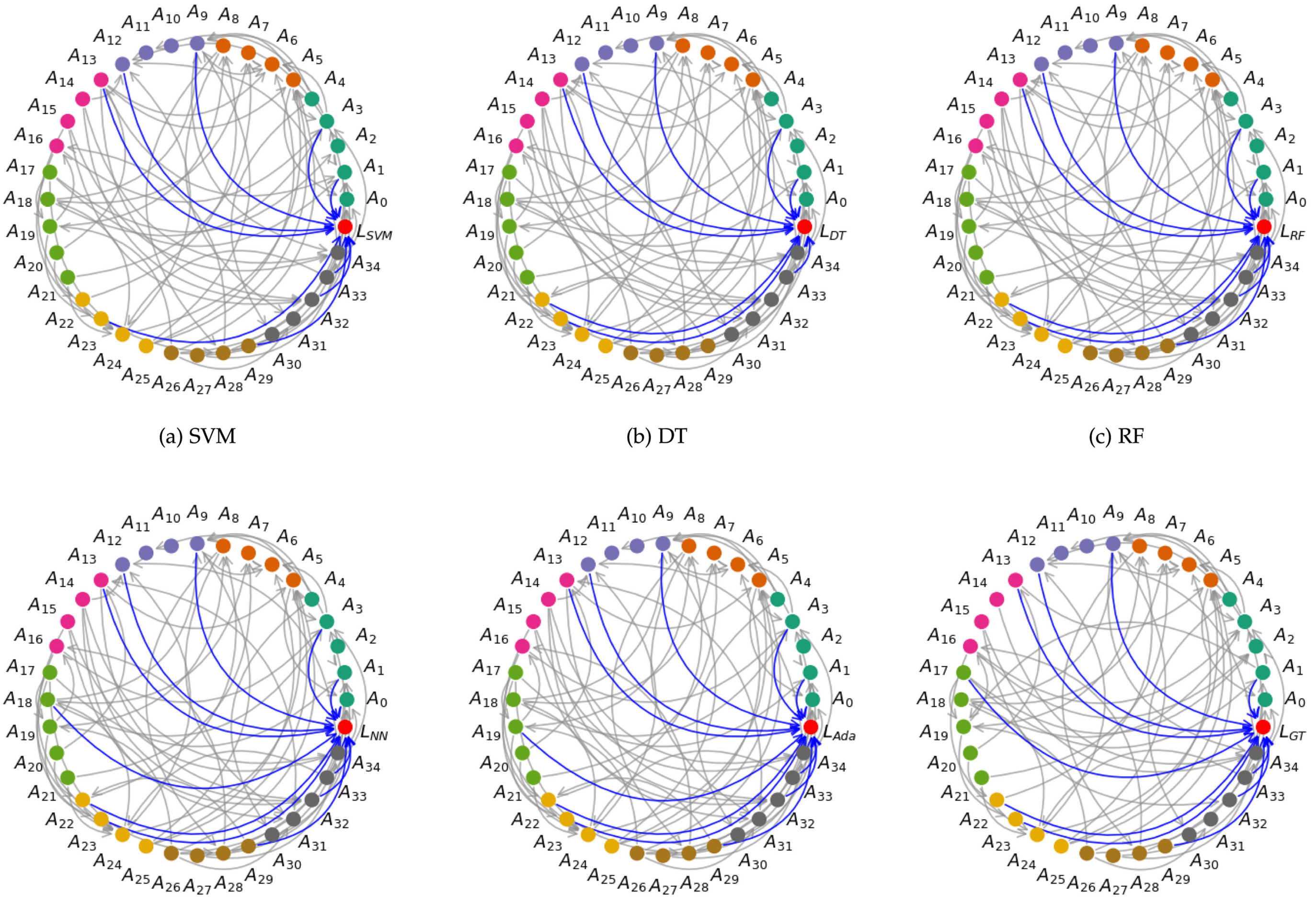
As none of the existing methods achieve both the two goals (1) automatic concept extraction, (2) concept-level black-box model interpretation, we qualitatively show the interpretation generated by our proposed CMIC framework for the following classifiers: SVM [6], Decision Tree (DT) [4], Random Forest (RF) [3], Neural Network (NN) [16], and AdaBoost (Ada) [12]. Those classifiers' classification quality is listed in Table 3. From the table, in the Bank Marketing dataset, the adopted classifiers all have low-performance scores, while most of the classifiers work well in the Divorce dataset. In the rest of this subsection, this phenomenon's explanations will be provided by using the interpretations generated by our CMIC framework.

### 5.2.3 Result Analysis

Figs. 5 and 6 show the generated causal graphs of different classifiers on two datasets. The ground truth label causal graph is obtained by running the SELF algorithm, mentioned in Section 4.2, on the original data. The node marked as $L$ (the red node) in each sub-figure denotes the classifier's output label, and the node named as $A_i$ is the $i$-th extracted concept. The edges between the direct cause of $f$-label and $f$-label $L_f$ are marked as blue.

From the figure, the causal graphs of different classifiers vary, which, to a certain degree, explains why the classifiers have low classification quality. Compared with the causal graph of the ground-truth label, Figs. 5f and 6c, none of the classifiers embed all relevant concepts; even worse, some classifiers predict the label based on some irrelevant concepts.

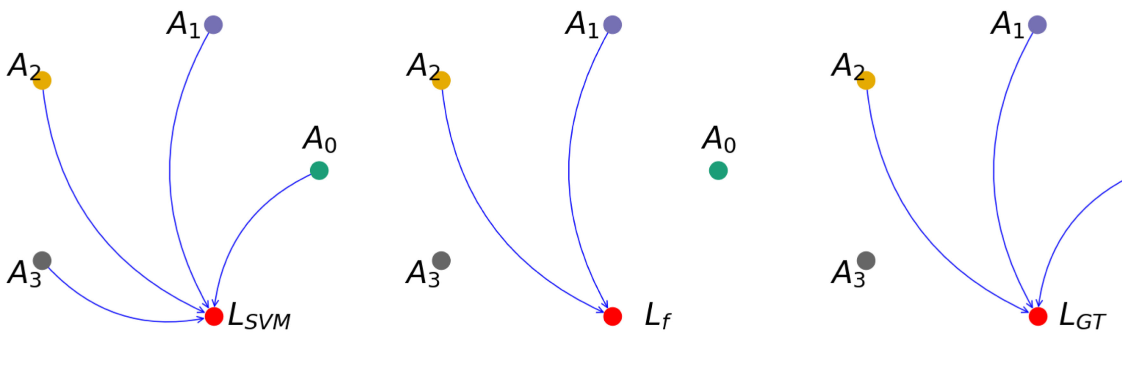
We also list the execution time of CMIC on two datasets in Table 4. From the table, it can be observed that it takes




(a) SVM (b) DT (c) RF

(d) NN (e) Ada (f) Groundtruth Label

Fig. 5. Causal graph results on bank marketing dataset.




(a) SVM (b) DT, RF, NN, Ada (c) Groundtruth Label

Fig. 6. Causal graph results on divorce dataset.

more time on the Bank Marketing dataset. The reason is that the Bank Marketing dataset is much larger than the Divorce dataset, which leads to more time consumed in the concept extraction component.

Besides the causal graph, the effects of relevant concepts associated with each classifier are shown in Table 5. The NA indicates that the concept is not relevant. The positive effect means the appearance of this concept would increase the probability of the label being positive, and the negative effect leads in the opposite direction. It is observed that classifiers with high prediction quality have similar concept effects with the ground truth.

We also conduct an additional experiment to validate the extracted concepts by applying intervention to the original data. Specifically, the concept of the one data record is flipped by changing the corresponding features within/out of the concept meaning scope. Based on the modified data, the classifiers then make the prediction. If the concept is

TABLE 4
Running Time

| | Bank Marketing | Divorce |
|---|---|---|
| Running Time (s) | 149.96 | 10.21 |

indeed relevant and its extract meaning is meaningful, the prediction will differ from the original one. Table 6 shows the change portion of the predicted label when intervention on the most relevant concepts. From the table, when flipping the concept, most of the classifier prediction changes, indicating the relevance of the concepts.

#### 5.2.4 Concept Meaning

Tables 7 and 8 list the detailed meanings of the concepts that are the direct cause of the $f$-label on the Bank Marketing dataset and Divorce dataset, respectively. In Table 8, $Q_i$ denotes the $i$-th question, and the following interval represents the range of its answers in the concept. For example, $Q_1$ is "If one of us apologizes when our discussion deteriorates, the discussion ends", and $Q_2$ is "I know we can ignore our differences, even if things get hard sometimes". More details of each are available on the data source page.

From the previous concept effect table, Table 5, concept $A_9$ is the most important concept, which positively affects the label. By checking the concept meaning in Table 7, it indicates that married aged people, who don't have housing loans and didn't receive the previous marketing campaign, tend to subscribe to the bank product, if the duration of their last contact is long. This case coincides with our common sense: Married old people without housing loans usually have generous savings or pensions, and the long duration of the last contact indicates their willingness to subscribe. The results validate that our CMIC framework is able to extract high-quality concepts and provide reasonable model interpretations.

TABLE 6
Label Prediction Change Portion

| Dataset | Concept | SVM | DT | RF | NN | Ada |
|---|---|---|---|---|---|---|
| Bank | $A_0$ | 9.5% | 24.1% | 15.2% | 7.4% | 6.6% |
| | $A_3$ | 10.1% | 20.1% | 10% | 8.1% | 4.1% |
| | $A_9$ | 9.5% | 46.6% | 38.1% | 18.2% | 31.8% |
| Divorce | $A_1$ | 38.2% | 21.8% | 0.6% | 1.2% | 0.6% |
| | $A_2$ | 38.2% | 45.3% | 37.6% | 29.4% | 40.0% |

## 6 Conclusion

Interpreting machine learning models at the concept level assists in providing more understandable reasoning of the model prediction. In this work, we propose the CMIC framework, which automatically extracts meaningful concepts, and discovers the causal relations between the concepts and model predicted labels to explain the model prediction. In the proposed CMIC framework, the concepts which serve as the cause of the model predicted label contribute to the model prediction, and the causal effects indicate their importance in model prediction. In the experiments, we quantitatively and qualitatively evaluate the extracted concepts as well as the generated interpretation using our CMIC framework. Results show that CMIC can generate meaningful concept-level model interpretations, which could also explain the behaviors of different classifiers.

*Future Work.* In this work, we focus on automatically extracting the meaningful concepts and analyzing the effect of the concepts on the prediction in a post-hoc manner. There are some future directions: (1) As the learned concept-based explanations, to some degree, indicate the reasons for different performances, utilizing the learned explanations to improve machine learning training is one of the future directions. (2) Concept extraction is the basis of our framework, therefore how to extract more human-

TABLE 5
The Effect of the Concept to the Prediction

| Dataset | Concept | SVM | DT | RF | NN | Ada | GT |
|---|---|---|---|---|---|---|---|
| Bank | $A_0$ | −0.10 | −0.24 | −0.18 | −0.09 | −0.08 | −0.12 |
| | $A_1$ | 0.07 | 0.12 | 0.14 | 0.12 | 0.11 | 0.13 |
| | $A_3$ | 0.21 | 0.14 | 0.12 | 0.01 | 0.11 | NA |
| | $A_9$ | 0.23 | 0.25 | 0.27 | 0.30 | 0.32 | 0.24 |
| | $A_{12}$ | −0.09 | −0.23 | −0.17 | −0.09 | −0.08 | −0.11 |
| | $A_{13}$ | 0.06 | 0.07 | 0.10 | 0.04 | 0.05 | 0.06 |
| | $A_{17}$ | NA | NA | NA | NA | NA | 0.01 |
| | $A_{18}$ | NA | NA | NA | −0.09 | NA | NA |
| | $A_{19}$ | NA | NA | NA | NA | 0.13 | NA |
| | $A_{22}$ | NA | −0.49 | −0.25 | −0.17 | −0.15 | −0.28 |
| | $A_{23}$ | −0.05 | −0.09 | −0.08 | −0.08 | −0.08 | −0.08 |
| | $A_{29}$ | −0.11 | NA | −0.23 | −0.11 | −0.10 | −0.15 |
| | $A_{32}$ | −0.02 | NA | −0.04 | −0.07 | −0.06 | −0.04 |
| | $A_{33}$ | 0.01 | −0.01 | −0.01 | −0.06 | −0.03 | −0.02 |
| | $A_{34}$ | −0.08 | −0.13 | NA | −0.16 | −0.08 | −0.06 |
| Divorce | $A_0$ | −0.61 | NA | NA | NA | NA | 0.51 |
| | $A_1$ | −0.55 | −0.62 | −0.71 | −0.71 | −0.71 | −0.72 |
| | $A_2$ | −0.56 | −0.65 | −0.73 | −0.73 | −0.71 | −0.72 |
| | $A_3$ | 0.48 | NA | NA | NA | NA | NA |

TABLE 7
Detected Concepts and Their Meanings on Bank Marketing Dataset

| Concept | Concept Meaning |
|---|---|
| $A_0$ | last contact duration: [23.25, 92.0];<br>number of contacts performed during this campaign and for this client: [14.5, 25.75];<br>number of days that passed by after the client was last contacted from a previous campaign: [-1];<br>number of contacts performed before this campaign and for this client: [0];<br>job admin.: 0; job entrepreneur: 0; job student: 0; education primary: 0;<br>outcome of the previous marketing campaign unknown: 1 |
| $A_1$ | age: [45, 71]; number of contacts performed during this campaign and for this client: [1.0, 2.0]<br>number of contacts performed before this campaign and for this client: [2.0, 9.0]<br>job housemaid: 0; marital single: 0; education unknown: 0;<br>outcome of the previous marketing campaign, success: 1 |
| $A_3$ | age: [25, 36]; average yearly balance: [3.75, 683.5]; job admin.: 0; job management: 0;<br>job retired: 0; job self-employed: 0; job student: 0; job technician: 0; marital single: 1;<br>outcome of the previous marketing campaign failure: 0 |
| $A_9$ | age: [65, 79]; has housing loan?: 0; last contact duration: [206.0, 503.25];<br>marital single: 0; education tertiary: 0;<br>outcome of the previous marketing campaign, failure: 1; |
| $A_{12}$ | has personal loan?: 1; last contact duration campaign: [22.5, 93.0] ;<br>number of contacts performed during this campaign and for this client pdays: [14.75, 25.0];<br>number of days that passed by after the client was last contacted from a previous campaign previous: -1;<br>number of contacts performed before this campaign and for this client: 0 |
| $A_{13}$ | has credit in default?: 0; has personal loan?: 0; job management:0; job entrepreneur: 0; education primary: 0;<br>number of contacts performed during this campaign and for this client: [1.0, 1.25];<br>contact communication type is cellular: 1 |
| $A_{17}$ | has housing loan?: 0; last contact duration: [266.25, 961.75] ;<br>number of contacts performed during this campaign and for this client: [1.0, 2.0];<br>number of contacts performed before this campaign and for this client: [0.0, 2.25];<br>job admin.: 0; job management: 0; job self-employed: 0; marital divorced: 0;<br>outcome of the previous marketing campaign, other: 0; |
| $A_{18}$ | average yearly balance: [119.75, 821.75]; job retired: 0;<br>job services: 0; education tertiary: 1;<br>outcome of the previous marketing campaign: success: 0 |
| $A_{19}$ | number of contacts performed before this campaign and for this client: [7.75, 18.25]<br>job student: 0; outcome of the previous marketing campaign is success: 0; |
| $A_{22}$ | last contact duration: [59.5, 265.75]; marital divorced: 1; education primary: 0;<br>contact communication type is telephone: 0; contact communication type is unknown: 0;<br>number of days that passed by after the client was last contacted from a previous campaign: -1; |
| $A_{23}$ | age: [32, 47]; number of contacts performed during this campaign and for this client: [8.5, 21.75];<br>number of days that passed by after the client was last contacted from a previous campaign previous: -1;<br>number of contacts performed before this campaign and for this client: 0; job admin.: 0;<br>outcome of the previous marketing campaign unknown: 0 |
| $A_{29}$ | last contact duration: [27.0, 93.0]; job entrepreneur: 0; marital divorced: 0;<br>number of contacts performed during this campaign and for this client: [3.0, 25.25]; |
| $A_{32}$ | last contact duration: [148.5, 281.75];<br>number of contacts performed during this campaign and for this client: 1;<br>job blue collar: 0; job entrepreneur: 0; job housemaid: 0; job management: 0;<br>marital divorced: 0; outcome of the previous marketing campaign, failure: 1 |
| $A_{33}$ | has personal loan?: 0; job retired: 0; education primary: 0; education primary: 0; job unemployed: 0;<br>outcome of the previous marketing campaign is failure: 0 |
| $A_{34}$ | number of contacts performed before this campaign and for this client: [3.75, 10.5]; job blue-collar: 0;<br>job management: 0;contact communication type is telephone: 0;<br>outcome of the previous marketing campaign is success: 0; |

TABLE 8
Detected Concepts and Their Meanings on Divorce Dataset

| Concept | Concept Meaning |
|---|---|
| $A_0$ | Q1: [0.0, 0.0]; Q8: [0.0, 0.0]; Q21: [0.0, 0.0]; Q22: [0.0, 0.0]; Q28: [0.0, 0.0]; Q29: [0.0, 0.0] ; Q30: [0.0, 0.0]; Q35: [0.0, 0.0] ; Q36: [0.0, 0.0]; Q38: [0.0, 1.0]; Q44: [0.0, 0.0]; Q46: [0.75, 3.0] ; Q54: [0.0, 1.0]; |
| $A_1$ | Q1: [0.0, 0.0] ; Q8: [0.0, 0.0]; Q9: [0.0, 0.0]; Q35: [0.0, 0.0]; Q40: [0.0, 0.0] ; Q52: [0.0, 2.0]; Q54: [0.0, 0.0] |
| $A_2$ | Q3, Q10, Q11, Q13, Q14, Q15, Q16, Q18, Q19, Q20, Q23, Q24, Q32, Q33, Q37, Q53: [0.0, 1.0] Q4, Q5, Q7, Q8, Q21, Q22, Q25, Q26, Q27, Q28, Q29, Q17, Q34, Q35, Q43: [0.0, 0.0]; Q0 , Q9, Q30, Q39, Q31, Q40, Q48: [0.0, 2.0]; Q12, Q36, Q38, Q41, Q42, Q44, Q49, Q50: [1.0, 2.0]; Q46: [0.0, 3.0]; Q47: [2.0, 3.0]; |
| $A_3$ | Q1: [3.0, 3.0]; Q2: [2.0, 2.25]; Q3: [2.0, 3.0]; Q4: [2.0, 2.0]; Q5: [2.75, 3.0]; Q6: [2.0, 2.0]; Q7: [2.0, 2.25]; Q8: [3.0, 3.0]; Q10: [2.0, 2.0]; Q13: [3.0, 3.0]; Q15: [3.0, 3.0]; Q12: [2.0, 2.25]; Q14: [2.0, 2.0]; Q16: [2.0, 2.0]; Q17: [2.0, 3.0]; Q18: [3.0, 3.0]; Q22: [2.0, 3.0]; Q23: [1.75, 2.0] Q25: [2.0, 2.25]; Q28: [2.0, 2.0]; Q31: [3.0, 4.0]; Q32: [3.75, 4.0]; Q33: [3.0, 4.0]; Q34: [4.0, 4.0]; Q35: [3.75, 4.0]; Q36: [4.0, 4.0]; Q37: [3.75, 4.0]; Q38: [4.0, 4.0]; Q40: [4.0, 4.0]; Q41: [4.0, 4.0]; Q43: [4.0, 4.0]; Q48: [3.75, 4.0]; Q54: [3.75, 4.0]; |

readable concepts is still the future direction. (3) In this work, the explanations are interpreted in a global view. Another future direction is to generate easy-to-understand explanations for a single sample from the causal aspect.

## References


[1] A. R. Akula, S. Wang, and S.-C. Zhu, "CoCoX: Generating conceptual and counterfactual explanations via fault-lines," in *Proc. 33rd AAAI Conf. Artif. Intell.*, 2020, pp. 2594–2601.

[2] M. Ancona, C. Öztireli, and M. Gross, "Explaining deep neural networks with a polynomial time algorithm for Shapley values approximation," 2019, *arXiv:1903.10992*.

[3] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[4] L. Breiman, *Classification and Regression Trees*. Evanston, IL, USA: Routledge, 2017.

[5] R. Cai, J. Qiao, Z. Zhang, and Z. Hao, "SELF: Structural equational likelihood framework for causal discovery," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1787–1794.

[6] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

[7] J. Chen, L. Song, M. J. Wainwright, and M. I. Jordan, "L-Shapley and C-Shapley: Efficient model interpretation for structured data," 2018, *arXiv:1808.02610*.

[8] M. Cuturi and A. Doucet, "Fast computation of Wasserstein barycenters," in *Proc. 31th Int. Conf. Mach. Learn.*, 2014, pp. 685–693.

[9] A. Datta, S. Sen, and Y. Zick, "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems," in *Proc. IEEE Symp. Secur. Privacy*, 2016, pp. 598–617.

[10] I. S. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2001, pp. 269–274.

[11] M. Du, N. Liu, Q. Song, and X. Hu, "Towards explanation of DNN-based prediction with guided feature inversion," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 1358–1367.

[12] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *Proc. Eur. Conf. Comput. Learn. Theory*, 1995, pp. 23–37.

[13] X. Gao, T. Mu, J. Y. Goulermas, J. Thiyagalingam, and M. Wang, "An interpretable deep architecture for similarity learning built upon hierarchical concepts," *IEEE Trans. Image Process.*, vol. 29, pp. 3911–3926, 2020.

[14] D. Georgiev, P. Barbiero, D. Kazhdan, P. Veličković, and P. Liò, "Algorithmic concept-based explainable reasoning," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 6685–6693.

[15] A. Ghorbani, J. Wexler, J. Zou, and B. Kim, "Towards automatic concept-based explanations," 2019, *arXiv:1902.03129*.

[16] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[17] Y. Goyal, U. Shalit, and B. Kim, "Explaining classifiers with causal concept effect (CaCE)," 2019, *arXiv:1907.07165*.

[18] F. Grün, C. Rupprecht, N. Navab, and F. Tombari, "A taxonomy and library for visualizing learned features in convolutional neural networks," 2016, *arXiv:1606.07757*.

[19] W. Guo, S. Huang, Y. Tao, X. Xing, and L. Lin, "Explaining deep learning models–A Bayesian non-parametric approach," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 4514–4524.

[20] M. Huai, J. Liu, C. Miao, L. Yao, and A. Zhang, "Towards automating model explanations with certified robustness guarantees," in *Proc. 36th AAAI Conf. Artif. Intell.*, 2022, pp. 6935–6943.

[21] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, no. 1, pp. 193–218, 1985.

[22] M. Ibrahim, M. Louie, C. Modarres, and J. Paisley, "Global explanations of neural networks: Mapping the landscape of predictions," 2019, *arXiv:1902.02384*.

[23] S. Jonnalagadda, T. Cohen, S. Wu, and G. Gonzalez, "Enhancing clinical concept extraction with distributional semantics," *J. Biomed. Informat.*, vol. 45, no. 1, pp. 129–140, 2012.

[24] B. Kim et al., "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," 2017, *arXiv:1711.11279*.

[25] Y. Kluger, R. Basri, J. T. Chang, and M. Gerstein, "Spectral biclustering of microarray data: Coclustering genes and conditions," *Genome Res.*, vol. 13, no. 4, pp. 703–716, 2003.

[26] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1885–1894.

[27] I. Kononenko et al., "An efficient explanation of individual classifications using game theory," *J. Mach. Learn. Res.*, vol. 11, no. Jan., pp. 1–18, 2010.

[28] H. Lakkaraju, S. H. Bach, and J. Leskovec, "Interpretable decision sets: A joint framework for description and prediction," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1675–1684.

[29] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4765–4774.

[30] D. Mincu et al., "Concept-based model explanations for electronic health records," in *Proc. Conf. Health Inference Learn.*, 2021, pp. 36–46.

[31] S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," *Decis. Support Syst.*, vol. 62, pp. 22–31, 2014.

[32] J. Pearl, *Causality*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[33] N. Puri, P. Gupta, P. Agarwal, S. Verma, and B. Krishnamurthy, "MAGIX: Model agnostic globally interpretable explanations," 2017, *arXiv:1706.07160*.

[34] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 97–101.

[35] A. Sarkar, D. Vijaykeerthy, A. Sarkar, and V. N. Balasubramanian, "A framework for learning Ante-hoc explainable models via concepts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10 286–10 295.

[36] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[37] I. Shpitser and J. Pearl, "Identification of joint interventional distributions in recursive semi-Markovian causal models," in *Proc. 21st Nat. Conf. Artif. Intell. 18th Innov. Appl. Artif. Intell. Conf.*, 2006, pp. 1219–1226.

[38] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 3145–3153.

[39] S. Singh, A. Gupta, and A. A. Efros, "Unsupervised discovery of mid-level discriminative patches," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 73–86.

[40] K. Sokol and P. A. Flach, "Counterfactual explanations of machine learning predictions: Opportunities and challenges for AI safety," *Proc. Workshop Artif. Intell. Saf. 33rd AAAI Conf. Artif. Intell.*, vol. 2301, 2019.

[41] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowl. Inf. Syst.*, vol. 41, no. 3, pp. 647–665, 2014.

[42] S. Tan, R. Caruana, G. Hooker, P. Koch, and A. Gordo, "Learning global additive explanations for neural nets using model distillation," 2018, *arXiv:1801.08640*.

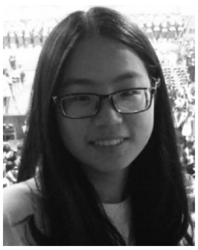
[43] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Amsterdam, The Netherlands: Elsevier, 2006.

[44] C. Villani, *Optimal Transport: Old and New.*, Berlin, Germany: Springer, 2008.

[45] C. Yang, A. Rangarajan, and S. Ranka, "Global model interpretation via recursive partitioning," 2018, *arXiv:1802.04253*.

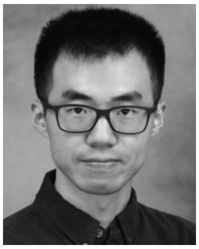
[46] C.-K. Yeh, B. Kim, S. O. Arik, C.-L. Li, P. Ravikumar, and T. Pfister, "On concept-based explanations in deep neural networks," 2019, *arXiv:1910.07969*.

[47] M. K. Yöntem, K. Adem, T. İlhan, and S. Kılıçarslan, "Divorce prediction using correlation based feature selection and artificial neural networks," *Nevşehir Hacı Bektaş Veli Üniversitesi SBE Dergisi*, vol. 9, no. 1, pp. 259–273, 2019.



**Liuyi Yao** received the BS degree in statistics from Nanjing University, in 2015, and the PhD degree from the Department of Computer Science and Engineering, SUNY Buffalo, in 2020. She is currently a research scientist with DAMO Academy, Alibaba Group. Her research interests include causal inference, time series analysis, and fairness.

**Yaliang Li** received the PhD degree from the Department of Computer Science and Engineering, SUNY Buffalo, in 2017. He is a research scientist with DAMO Academy, Alibaba Group. Before that, he worked as a research scientist with Baidu Research, and a senior researcher with Tencent Medical AI Lab. He is broadly interested in machine learning and data mining with a focus on truth discovery, knowledge graph, question answering, differential privacy, recommendation, and more recently automated machine learning.

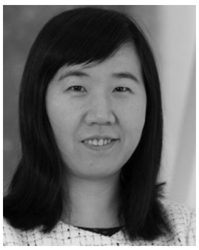


**Sheng Li** (Senior Member, IEEE) received the BEng degree in computer science and engineering and the MEng degree in information security from the Nanjing University of Posts and Telecommunications, China, in 2010 and 2012, and the PhD degree in computer engineering from Northeastern University, Boston, MA, in 2017. He is a Tenure-Track assistant professor with the School of Data Science, University of Virginia. Previously, he was an assistant professor with the Department of Computer Science, University of Georgia from 2018 to 2022, and was a data scientist with Adobe Research from 2017 to 2018. He has published more than 120 papers at peer-reviewed conferences and journals, and has received more than 10 research awards, such as the INNS Aharon Katzir Young Investigator Award, Adobe Data Science Research Award, SDM Best Paper Award, and IEEE FG Best Student Paper Honorable Mention Award. He has served as an associate editor for seven journals such as the *IEEE Transactions on Neural Networks and Learning Systems*, *IEEE Transactions on Circuits and Systems for Video Technology*, and *IEEE Computational Intelligence Magazine*. He has also served as area chair/senior program committee member for NeurIPS, ICLR, AAAI, IJCAI, SDM, and ICPR. His research interests include trustworthy representation learning, causal inference, visual intelligence, and user behavior modeling.

**Jinduo Liu** received the BS and PhD degrees in computer science and technology from the Beijing University of Technology, Beijing, China, in 2013 and 2020, respectively. He is currently a lecturer with the Beijing Artificial Intelligence Institute, Faculty of Information Technology, Beijing University of Technology. His current research interests include machine learning, data mining, and brain informatics.

**Mengdi Huai** is an assistant professor with the Department of Computer Science, Iowa State University. Her research interests include the general areas of data mining and machine learning, with an emphasis on developing novel techniques to build trustworthy learning systems that are explainable, robust, private, and fair.

**Aidong Zhang** (Fellow, IEEE) is a William Wulf faculty fellow and professor with the University of Virginia. Her research interests include machine learning, data mining/data science, bioinformatics, and health informatics. She has authored more than 380 research publications in these areas. She is a fellow of the ACM and AIMBE.



**Jing Gao** received the PhD degree from Computer Science Department, University of Illinois at Urbana Champaign, in 2011, and subsequently joined University at Buffalo, in 2012. She is currently an associate professor with the Elmore Family School of Electrical and Computer Engineering, Purdue University. Before joining Purdue in January 2021, she was an associate professor with the Department of Computer Science and Engineering, University at Buffalo (UB), State University of New York. She is broadly interested in data and information analysis with a focus on data mining. In particular, she is interested in information veracity analysis, crowdsourcing, knowledge graphs, multi-source data analysis, anomaly detection, transfer learning, text mining and data stream mining as well as various data mining applications in healthcare, bioinformatics, social science, transportation, cyber security, and education. She has published more than 150 papers in referred journals and conferences with more than 10,000 citations. She is a recipient of NSF CAREER Award, IBM faculty award, ICDM Tao Li Award and SDM/IBM Early Research Career Award.