scientific reports

Check for updates

OPEN

An interpretable model based on concept and argumentation for tabular data

Haixiao Chi[1]✉, Dawei Wang[2], Beishui Liao[3]✉, Gaojie Cui[2] & Feng Mao[2]

Interpretability has become an essential topic for artificial intelligence in some high-risk domains such as healthcare, banking, and security. For commonly used tabular data, traditional methods trained end-to-end machine learning models with numerical and categorical data only and did not leverage human-understandable knowledge such as data descriptions. Yet mining human-level knowledge from tabular data and using it for prediction remain a challenge. In this paper, we propose a novel component for tabular data, called quantitative argumentation layer, which mined concepts from both data and data descriptions. we construct a concept and argumentation model (CAM) that embeds human-aligned reasoning processes—quantitative argumentation explicitly represents domain knowledge through human-understandable argumentation rules rather than opaque machine encodings. As a result, CAM provides decisions that are based on human-level knowledge and the reasoning process is intrinsically interpretable. Finally, to explain the proposed interpretable model, we provide a dialogical explanation containing dominated reasoning paths within CAM. Human-subject evaluations indicate CAM is comprehensible to individuals, and the explanations provide reasonable rationales and have a high level of user acceptance. We also conduct data experiments on both open-source benchmarks and real-world business datasets that show that our interpretable approach can reach competitive results compared with state-of-the-art models.

Keywords Quantitative argumentation, Explainable AI, Risk assessment, Decision-making model

For decision-making tasks related to high-risk domains, machine learning (ML) methods are required to have a high level of interpretability[1,2]. Many post-hoc and feature-based explainable methods, such as SHapley Additive exPlanations (SHAP)[3] and local interpretable model-agnostic explanations (LIME)[4], have been proposed to explain black-box models. However, Slack et al. showed that post-hoc explanation can neither reflect the real behavior of a black-box model nor improve human understanding of the model[5]. Thus, developing interpretable models has been an increasingly active research direction. Many interpretable models were proposed such as explainable boosting machine (EBM)[6], neural generalized additive model (NODE-GAM)[7], neural basis models (NBM)[8], GRAND-SLAMIN[9], etc. These state-of-the-art models can provide a feature-based explanation for a prediction target. However, they are purely data-driven and do not include domain knowledge from experts, which may lead to different explanations for the same decision based on different fitting functions. So, such purely data-driven explainable models lack robustness and trustworthiness.

Alongside feature-based methods, in recent years argumentative XAI is increasingly showing benefits for various fields[10–13]. Argumentative XAI applies computational argumentation to extract argumentation frameworks (AFs) to explain the process and output of the underlying model. Whereas feature-based methods focus on the input–output behavior of the underlying model, AFs as explanations point to the dialectical relationships among arguments, abstractly representing interactions among the inner components of the underlying model. These frameworks offer a transparent and interactive way for users to understand a model's reasoning, which can help in debugging and improving the model's performance[14]. However, a major limitation of current approaches is their heavy reliance on a pre-defined argumentation structure. This structure is either manually defined by human experts[15] or derived from the inherent properties of the data, such as the tree structure in social media or polling data[14,16,17]. Such a dependency makes it challenging to apply these methods to tabular data, where explicit hierarchical structures are usually absent and relying on domain experts to manually construct the implicit tree-like knowledge is costly compared to automated approaches.

[1]Xiamen Medical College, Xiamen 361023, China. [2]Alibaba Group, Hangzhou 310028, China. [3]Zhejiang University, Hangzhou 310028, China. ✉email: haixiaochi@zju.edu.cn; baiseliao@zju.edu.cn

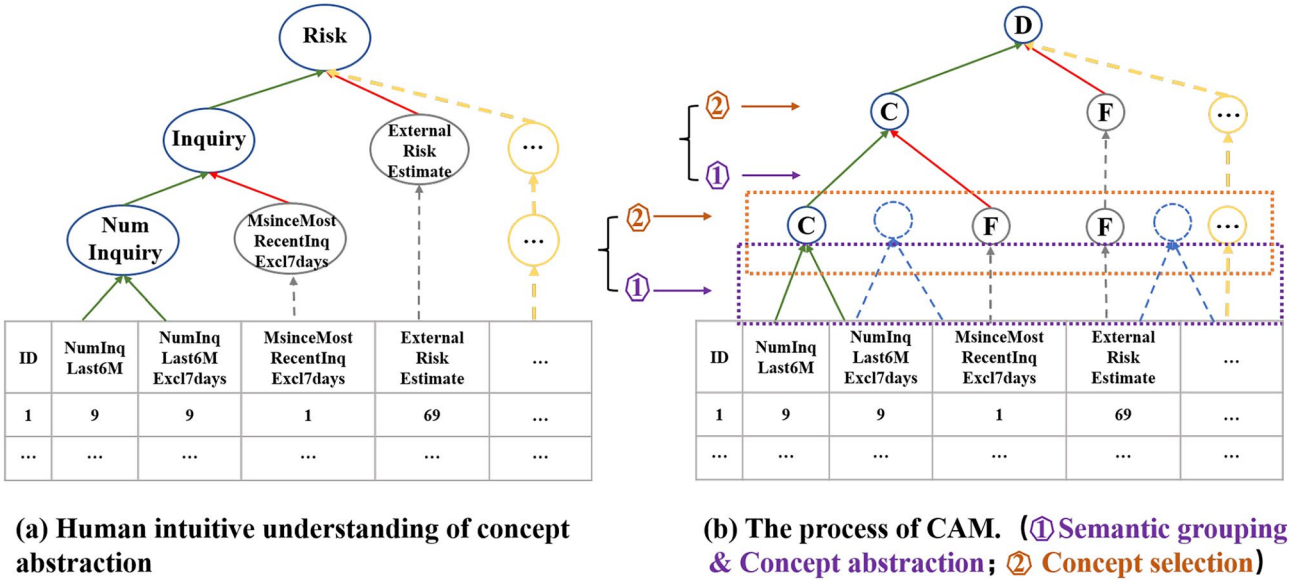


Fig. 1. A concrete example of risk estimation to illustrate our idea: (**a**) Human intuitive understanding of concept generation for risk assessment. The green and red lines represent support and attack relations from the lower nodes to the higher nodes, respectively. The yellow nodes represent concepts or features, and the yellow dashed lines indicate the omitted concept generation process. The grey dashed lines link features or concepts which are not grouped with others. (**b**) A P_{QAL} firstly groups features with similar semantics together to abstract concepts and secondly mines the support or attack relations, evaluates the concepts, and keeps the important ones. CAM organizes P_{QAL} to repeat this process until no concept can be generated. The concept of the decision node is linked on the top of the stacked QALs to form a QAF for risk assessment. C in the blue circle denotes the selected concept, while the blue dashed circles represent the deleted ones, F in the grey circle represents features, and D denotes the concept of the decision node. The blue dashed lines indicate the deleted candidates for semantic grouping.

To address these limitations, our work champions a fusion process of knowledge and data in the modeling process. We propose a model based on concept and argumentation (called CAM[1]) that automatically integrates knowledge with data by constructing argumentative structures from tabular data. In this paper, we define a "concept" as a higher-level semantic unit abstracted from a group of related, fine-grained features. While individual features in tabular data are already semantically interpretable (e.g., 'number of inquiries in the last 6 months'), they often represent specific measurements. CAM aims to mirror the human cognitive process of grouping these specific details into broader, more abstract ideas (e.g., grouping several inquiry-related features into a single concept of 'Inquiry'). This hierarchical abstraction provides a more concise and intuitive explanation of the model's reasoning.

In conventional data modeling approaches, textual descriptions accompanying tabular data are often overlooked. However, CAM leverages this human-comprehensible knowledge to perform concept mining, subsequently identifying relationships among these concepts within the data to construct an argumentative tree structure. This tree structure aligns with neural network architectures, enabling the utilization of neural network algorithms in model construction. Crucially, at the interpretative level, this approach allows for the complete elucidation of the traditionally "black box" neural network model, whereby each neuron and its interconnections find a corresponding, human-interpretable knowledge representation within the argumentative tree.

The CAM's modeling process uniquely combines knowledge-guided and data-driven methodologies. This synthesis results in a model that exhibits high levels of accuracy while maintaining transparency throughout its operational framework. By bridging the gap between machine learning techniques and human-understandable representations, CAM offers a promising avenue for developing more interpretable and accountable AI systems in decision-making processes for tabular data. Figure 1a shows an argumentative structure manually abstracted from features in a real-world risk assessment task[18]. For example, *NumInqLast6M* and *NumInqLast6Mexcl7days* are grouped as *NumInquiry*, which together with *MsinceMostRecentInqExcl7days* forms a higher-level concept *Inquiry*. This process continues until the decision node *Risk* is linked to the highest-level concepts. Support or attack relations are derived from data descriptions.

In contrast, Fig. 1b illustrates how CAM simulates this abstraction. A quantitative argumentation procedure (P_{QAL}) groups features with similar descriptions, learns relations among nodes, and selects important concepts

[1]We use CAM as the acronym for our Concept and Argumentation Model. We acknowledge the potential overlap with Class Activation Maps, also abbreviated as CAM in computer vision research. Our work, however, is situated in the distinct domain of tabular data and argumentation-based XAI.

for further abstraction. The process stops when no new concepts emerge, and the decision node is linked to the final structure, forming a quantitative argumentation framework (QAF). To achieve this, CAM integrates semantic knowledge mining with clustering and abstraction. A field-wise learning algorithm then evaluates candidate concepts by their contribution to model performance. Through this process, CAM learns weighted support/attack relations and infers concept values directly from data. Based on these mechanisms, CAM can highlight the most relevant reasoning paths and present them in dialogue-based natural language. Finally, we validate CAM from two perspectives: (1) human-centered studies on interpretability, coherence, and user acceptance, and (2) technical benchmarking on standard datasets and high-risk business scenarios, ensuring interpretability does not compromise predictive accuracy.

The remainder of this paper is organized as follows. In "Preliminary: quantitative argumentation" section provides preliminary knowledge on quantitative argumentation. In "An interpretable model based on concept and argumentation for tabular data" section introduces our interpretable model based on concept and argumentation, CAM, and details its decision-making process and method for generating explanations for tabular data. In "Experiments" section presents the results of the data experiments and user study, analyzing CAM's performance in terms of accuracy and interpretability. In "Related work" section reviews related work. Finally, in "Conclusions" concludes the paper, and in "Future work" section discusses the limitations of the current work and details future research directions

Preliminary: quantitative argumentation

Argumentation theory provides a natural framework for modeling reasoning under conflict and uncertainty, mirroring human dialectical processes of weighing pros and cons. This makes it a powerful candidate for building explainable AI systems that reason in a human-aligned manner. We now briefly introduce the core definitions of the quantitative argumentation framework (QAF), which serves as the mathematical foundation for CAM's structure and reasoning mechanism.

QAF involves not only logical knowledge such as arguments, attack, and support relations, but also numerical knowledge that can reflect the uncertainty of arguments and their relations. QAF is based on bipolar argumentation framework (BAF)[19] by quantifying the semantics of arguments and the relations between them. In this paper, we adopt QAF to represent bipolar argumentation framework[19] with quantitative arguments and relations, which can also be noted as the optimized quantitative argumentation debate (O-QuAD) framework[15] or edge-weighted quantitative bipolar argumentation frameworks (Edge-weighted QBAF)[20]. Many reasoning methods have been proposed for evaluating their semantics, including the DF-QuAD algorithm[21], the O-QuAD algorithm[15], the multi-layer perception (MLP)-based algorithm[20], etc.

Definition of QAF

A quantitative argumentation framework (QAF) is a formal model to represent arguments and their relationships. It is defined as a quadruple $\langle \mathscr{A}, E, \beta, \omega \rangle$, where:

- $\mathscr{A}$ is a set of arguments;
- $E \subseteq \mathscr{A} \times \mathscr{A}$ is a set of directed edges between arguments, which are acyclic;
- $\beta : \mathscr{A} \to [0, 1]$ assigns a *base score* $\beta(a)$ to each argument $a \in \mathscr{A}$;
- $\omega : E \to \mathbb{R}$ assigns a *weight* to each edge, indicating the strength and polarity of the relation.

For each argument $a \in \mathscr{A}$, we define its set of *attackers* and *supporters* as follows:

$$Att(a) = \{b \in \mathscr{A} \mid (b, a) \in E \text{ and } \omega(b, a) < 0\},$$
$$Sup(a) = \{b \in \mathscr{A} \mid (b, a) \in E \text{ and } \omega(b, a) \geq 0\}.$$

Here, *Att*(*a*) denotes the set of arguments that negatively affect *a*, while *Sup*(*a*) denotes the set of arguments that positively affect *a*. In short, a QAF can be regarded as a graph where each argument has an initial score, and directed edges represent either supportive (positive weight) or attacking (negative weight) influences.

For example, Fig. 2 (left) illustrates part of the decision problem in Fig. 1. The decision node, *Risk*, represents the applicant's overall credit risk. It is supported by *A3: Inquiry*, which has two sub-nodes: *A1: NumInquiry* (recent inquiries) and *A2: MsinceMostRecentInqExcl7days* (months since the most recent inquiry excluding 7 days). More inquiries indicate higher risk, while longer periods without inquiries are favorable. *A4: ExternalRiskEstimate* attacks *Risk*; higher values reinforce confidence in creditworthiness. This decision problem can be modeled as a QAF:$\langle \mathscr{A} = \{Risk, A1, A2, A3, A4\}, E = \{(A3, Risk), (A4, Risk), (A1, A3), (A2, A3)\}, \beta, \omega \rangle$, with $Att(Risk) = \{A4\}, Sup(Risk) = \{A3\}, Att(A3) = \{A2\}, Sup(A3) = \{A1\}$.

The selection of reasoning method within QAF

Research has demonstrated a correspondence between multilayer perceptrons (MLPs) and QAF[20]. By translating the forward propagation mechanics of MLP as the reasoning method within QAF, it is feasible to apply MLP algorithms to obtain the quantitative knowledge of a QAF from tabular data automatically. Thus, the MLP-based reasoning method is selected as the reasoning algorithm within QAF.

In the MLP-based reasoning method of QAF, we have a strength $s(a) \in [0, 1]$. *s*(*a*) is the strength value of argument *a*. The strength values are then updated by doing the following two steps for all $a \in \mathscr{A}$ from bottom to top:

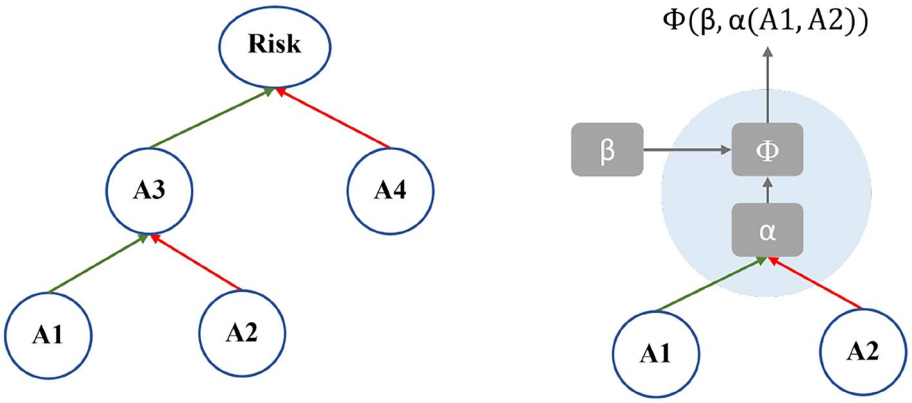


Fig. 2. Example of a QAF (left) and illustration of MLP-based reasoning process of a single modular (right).

$$\text{Aggregation: } \alpha(a) := \sum_{(b,a)\in E} \omega(b,a) \times s(b) \tag{1}$$

$$\text{Combination: } s(a) := \Phi\left(ln\left(\frac{\beta(a)}{1-\beta(a)}\right) + \alpha(a)\right) \tag{2}$$

where $\Phi(z) = \frac{1}{1+exp(-z)}$ is the logistic function. Let $ln(0) := -\infty$, $ln(\frac{1}{0} := \infty)$, $\Phi(-\infty) = 0$, $\Phi(\infty) = 1$ and for all $x \in \mathbf{R}, x - \infty = -\infty$ and $x + \infty = \infty$. In this way, the composition of the aggregation and combination function is continuous and always returns values from the closed interval [0, 1].

Each argument is initially assigned a strength using its base score function β. The strength values are then updated through an aggregation function α and a combination function Φ, as shown in Fig. 2 (right). The aggregation function α combines the strength values of attackers and supporters. The combination function Φ then integrates this aggregate with the base score to compute a new strength within the target domain. Intuitively, supporters increase the strength, while attackers decrease it.

An interpretable model based on concept and argumentation for tabular data

Knowledge representation

Knowledge in tabular data

In tabular data, the knowledge may be divided into two categories: human-level knowledge presented in data description and implicit knowledge learned from data. The former expresses the semantics of features in natural language and is intuitively understandable to humans. From a cognitive perspective[22], we believe that a concept should be the common characteristics abstracted from a set of features or lower-level concepts, which should be consistent with human cognition, as shown in Fig. 1. Through hierarchical abstraction, we can naturally obtain a tree structure of features and concepts that are understandable to humans, with a comprehensiveness level ranging from low to high. We denote this tree as the concept tree. Another kind of knowledge can be learned from data and expressed as the value of higher-level concepts relevant to the prediction targets which are aggregated by exploiting the correlative features or lower-level concepts.

Suppose $\mathscr{T} = \langle L, X, D\rangle$ is one type of tabular data structure, where L are descriptions of features, X is the data, and D is the decision target. In a tabular dataset of the $\mathscr{T}$ type, an instance $x \in \mathbb{R}^n$ in X is defined as n-element vector representing n scalar raw features in $\mathscr{F}$, where $\mathscr{F}$ is a set of the raw features contained in tabular data. In this paper, we assume that there are some underlying feature groups in a tabular data structure. The features in a group are semantically similar and target-relevant and can be abstracted for a more general semantic unit as a concept. Note that some features may not be in any group and some may be in multiple groups. We are interested in mining the concepts for a decision target from data description L, and data X, and utilizing quantitative argumentation for explicit knowledge representation and reasoning to form the interpretable decision-making model.

Representing knowledge in quantitative argumentation frameworks

A concept tree generated from a tabular data $\mathscr{T} = \langle L, X, D\rangle$ can be represented as a quantitative argumentation framework (QAF), denoted as $QAF_{\mathscr{T}} : \langle \mathscr{A}_{\mathscr{T}}, E_{\mathscr{T}}, \beta_{\mathscr{T}}, \omega_{\mathscr{T}}\rangle$. Here, each argument $a \in \mathscr{A}_{\mathscr{T}}$ represents a concept $c \in \mathscr{C}$ or a feature $f \in \mathscr{F}$, where $\mathscr{A}_{\mathscr{T}} = \mathscr{C} \cup \mathscr{F}$. The edges $E_{\mathscr{T}} \subseteq \mathscr{A}_{\mathscr{T}} \times \mathscr{A}_{\mathscr{T}}$ describe positive and negative correlations between these concepts and features. The framework also includes a function $\beta_{\mathscr{T}} : \mathscr{A}_{\mathscr{T}} \to [0, 1]$ that assigns a *base score* to each argument, and a function $\omega_{\mathscr{T}} : E_{\mathscr{T}} \to \mathbb{R}$ that assigns a weight to each edge.

In our QAF, the arguments representing features are at the leaf nodes. Their strength can be obtained directly from the data without the need for a base score function to assign initial values. Categorical features are target encoded, and all features are subsequently transformed using a quantile transformer with a uniform output

distribution, resulting in continuous values into a quantifiable strength score in the range [0, 1], which serves as the initial value for the feature arguments. The functions $\beta_{\mathscr{T}}$ and $\omega_{\mathscr{T}}$ for the remaining arguments will be defined when we introduce the field-wise learning algorithm.

Knowledge acquisition: procedure of quantitative argumentation layer

In this section, we utilize P_{QAL} to mine semantic and quantitative knowledge of the concepts as shown in Fig. 3. Semantic knowledge mining is realized by the semantic knowledge mining approach to automatically search lower-level knowledge units (such as features and lower-level concepts) with similar meanings and abstract higher-level concepts from the data description. Then, A field-wise learning algorithm is designed for quantitative knowledge mining by learning the values of concepts and their relations from data and evaluating the concepts for deleting the unimportant ones. In each P_{QAL}, the selected concepts and the ungrouped features can be represented as a QAL, and their semantics are noted in L' for the next P_{QAL}.

Unlike conventional clustering or summarization approaches that produce static semantic groupings, our P_{QAL} dynamically integrates semantic similarity with quantitative reasoning through the QAF structure. This coupling allows each cluster to be assigned interpretable argumentative roles (support/attack), learned adaptively via field-wise optimization, thereby transforming unsupervised grouping into a reasoning-guided semantic abstraction process.

Semantic knowledge mining

To simulate the process of abstracting concepts in human cognitive learning, we need to combine features with similar meanings and extract the same characteristics as the meaning of the generated concepts. To achieve this goal, natural language preprocessing is necessary for semantic knowledge mining[23]. A pretrained multi-lingual language model transfers the natural language information into a vector space. In that way, the meanings of features or concepts are embedded into vectors from natural sentences.

Given a data description L of tabular data $\mathscr{T}$, $l_a \in L$ represents the description of a feature a, and l_a is also a set of words such that description l_a can perform the intersection operation with other descriptions. It is worth noting that after the first round of P_{QAL}, the descriptions also contain descriptions of concepts, which means that a can represent a feature or a concept. After the embedding, the descriptions are transferred into vectors. We denote the vector version of L and l_a as $\hat{L}$ and $\hat{l_a}$ repectively. Suppose a group of features or concepts (denoted as $\mathscr{A}_c = \{a_j, \ldots, a_k\}, \mathscr{A}_c \subseteq \mathscr{A}_{\mathscr{T}}$) can be combined together to generate higher-level concept c. The

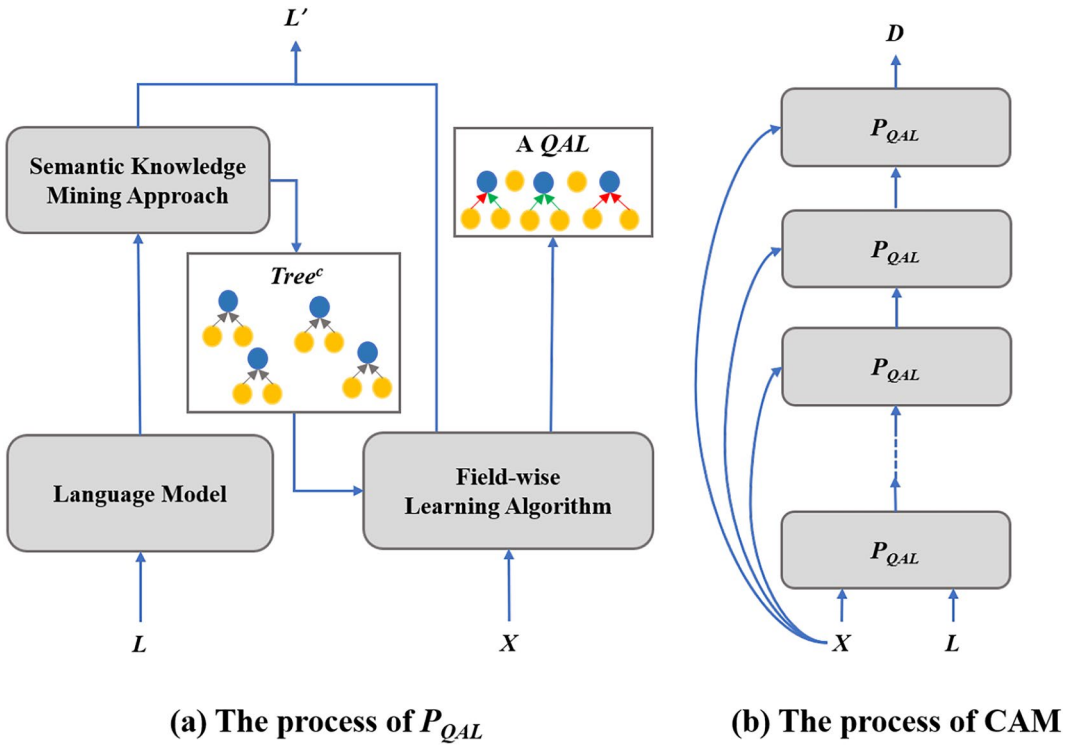


Fig. 3. The overall workflow of CAM. (**a**) Illustrating a P_{QAL}, which performs semantic grouping and concept abstraction by semantic knowledge mining approach, and performs concept selection by field-wise learning algorithm. Blue nodes represent newly generated concepts and yellow nodes represent features or lower-level concepts (whose children nodes are omitted in this figure). $Tree^c$ denotes a set of structures of the possible concepts. L' represents a new data description that contains the semantics of the ungrouped features (or concepts) and the newly generated concepts for the next P_{QAL}. (**b**) The architecture of CAM is built with stacked QALs by repeatedly utilizing P_{QAL}.

description of the generated concept can be defined as $l_c = \bigcap_{a \in \mathscr{A}_c} l_a$. The tree structure with root c is denoted as $tree_c = \{c : \{a_j, \ldots, a_k\}\}$. Our goal in semantic knowledge mining is to obtain all the possible concepts' descriptions L^c and structures $Tree^c$ from L, where L^c is a set of l_c and $Tree^c$ is a set of $tree_c$.

The clustering algorithm can be utilized to find the groups of descriptions. The agglomerative hierarchical clustering (AHC) algorithm[24] is adopted since it can capture the hierarchical relationship between clusters, which other clustering algorithms cannot achieve. The main idea of the AHC algorithm is that each object starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy to form the tree structure. The semantic knowledge mining approach is designed based on the AHC algorithm to conduct semantic grouping and concept abstraction, as described in Algorithm 1. In Algorithm 1, we have designed a threshold, denoted as μ, to ensure that the semantics within each cluster are sufficiently similar, thereby guaranteeing that the concepts generated within each cluster are meaningful.

Input: $L = \{l_1, ..., l_n\}$, $\hat{L} = \{\hat{l_1}, ..., \hat{l_n}\}$
Output: L^c, $Tree^c$
calculate the similarity matrix S, in which S_{jk} represents the cosine similarity of $\hat{l_j}$ and $\hat{l_k}$.
for *each* $l_j \in L$ **do**
  find the most similar semantics l_k by traversing the S.
  if $S_{jk} \geq \mu$ **then**
    group l_j and l_k together as a cluster $\{l_j, l_k\}$.
    abstract new possible concept c_i with semantics $l_{c_i} = l_j \bigcap l_k$.
    $L^c \Leftarrow l_{c_i}$.
    $Tree^c \Leftarrow \{c_i : \{a_j, a_k\}\}$.
  end
end

Algorithm 1. Semantic knowledge mining approach in P_{QAL}

Quantitative knowledge mining: the field-wise learning algorithm

After the process of semantic knowledge mining, assume that we obtain t possible concepts from n features and their structural information $Tree^c$. In this part, the quantitative knowledge of concepts from the data needs to be mined. The importance of concepts should be evaluated in order to eliminate the unimportant ones. Given training data X_{TR}, we split it into a sub-training set X_{tr} and a validation set $Xvld$, as shown in Fig. 4. Then we represent the knowledge in X_{tr} as a *QAF*, and with learning algorithm $\mathscr{L}$ learn a CAM model $\mathscr{L}(X_{tr}, QAF)$. To evaluate this CAM model, we use the validation set X_{vld} to evaluate the same *QAF* and calculate the performance $\mathscr{E}(\mathscr{L}(X_{tr}, QAF), X_{vld}, QAF)$, denoted as $\mathscr{E}(QAF)$ for short. The performance can result in metrics such as area-under-curve (AUC).

Many researchers[25] believe that the concept is "important" for the decision targets if its presence is necessary. Thus, we define the rule for evaluating the importance of concepts and select the important ones as follows:

$$\text{If } \mathscr{E}(QAF_i) \geq \mathscr{E}(QAF_0), \text{then keep } c_i. \text{ Else drop } c_i. \tag{3}$$

where $\mathscr{E}(QAF_0)$ represents the performance of original quantitative argumentation framework (denoted as QAF_0), $\mathscr{E}(QAF_i)$ represents the performance of the QAF_i, and QAF_i is the QAF by adding a new concept c_i to QAF_0.

Though highly accurate, direct evaluation for feature (and concept) sets is often rather expensive. In real-world business scenarios, training a model to converge may take great computational resources. Such direct evaluations are often too expensive to be invoked repetitively in the concept generation procedure. In order to improve the evaluation efficiency, we proposed a field-wise learning algorithm in CAM.

To accelerate concept evaluation, the field-wise learning algorithm runs in two steps. In the first step, we use X_{tr} to train a MLP for learning the strength of nodes and edges of QAF_0 from a previous QAL to concepts of decision node D and evaluate QAF_0 as $\mathscr{E}(QAF_0)$ on the validation set X_{vld}. We chose MLP as the learning model since QAF an MLP correspond in structure and reasoning approach.

In the second step, we link the newly generated concept c_i with its structural information $tree_c$ as a sub-framework of QAF_0 and delete the edges that link the children of c_i directly to D, thus we get a new QAF_i by adding $tree_c$ and removing the repeated arguments. Then, we use a MLP that has the same structure with QAF_i to learn the unknown strength of nodes and edges of QAF_i. The same parts of QAF_i and QAF_0 have been learned in the first step, thus the MLP only learns the strength of edges and nodes related to c_i. Hence, the learning process is 'field-wise' and can be processed in parallel.

Formally, the previous QAL is denoted as $\mathscr{A}_0 = \{a_1, \ldots, a_q\}$, where a_i, $1 \leq i \leq q$, may be features or concepts, and the value of a_i is denoted as $s(a_i)$. Especially, in the first round of concept mining, a_i only represents features, and $s(a_i)$ is the value in the interval of 0 to 1 obtained by data pre-processing of the feature value x_{a_i}. The structure of a newly generated concept c_i are denoted as $\{c_i : \{a_j, a_k\}\}$. And when a_i is a concept,

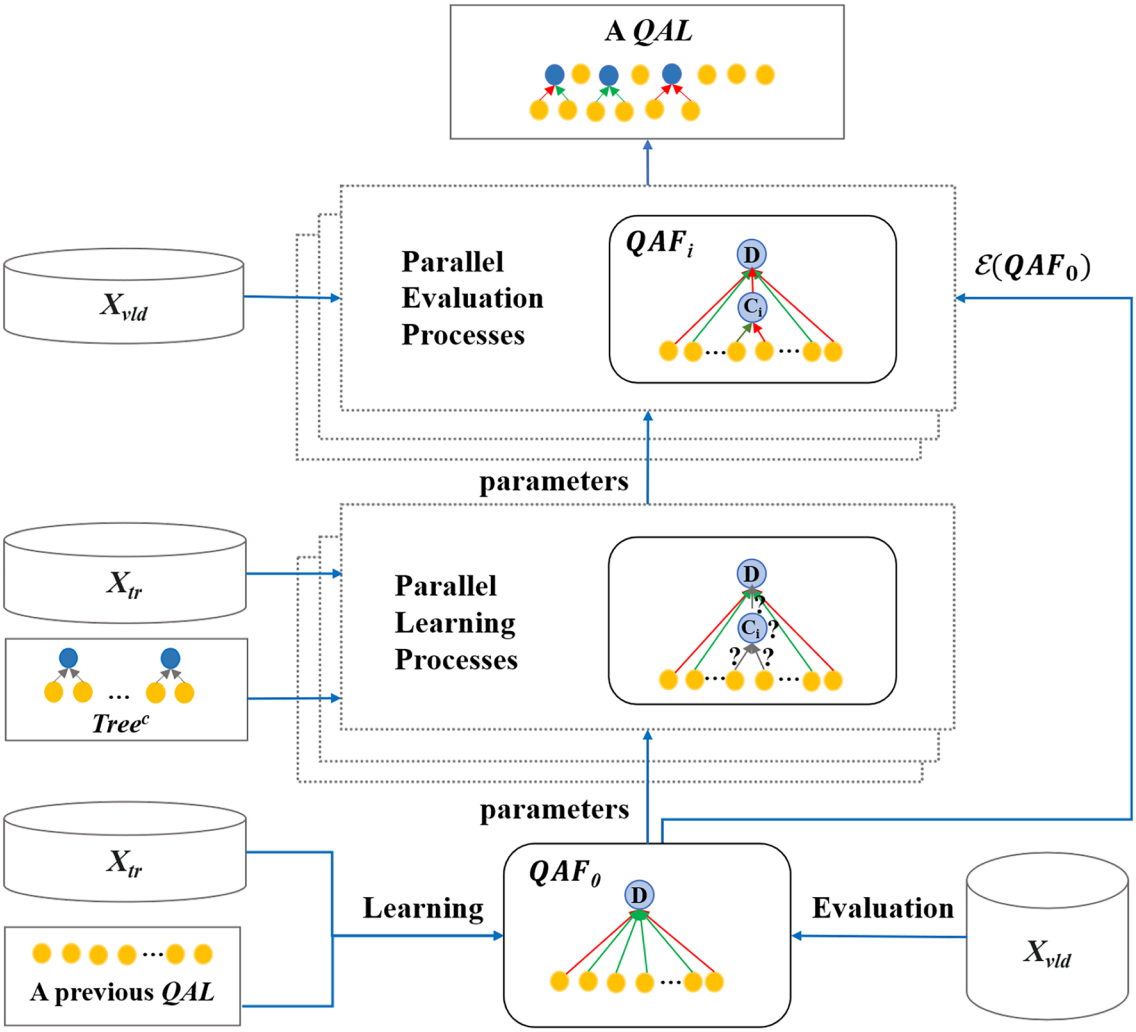


Fig. 4. Illustration of the field-wise learning algorithm for concept selection. The field-wise learning algorithm runs in two steps. In the first step, we learn and evaluate the QAF_0. And then in step 2, we conduct a parallel learning and evaluation process to assess the importance of each concept and select the important ones. Parameters represent the learned values of edges and nodes in QAF by MLP.

the quantitative information of a_i and its children is learned in the previous P_{QAL}. In the first step, the MLP can be described as:

$$s(D) = \Phi\left(\sum_{a_i \in \mathscr{A}_0} w_i \times s(a_i) + b_D\right) \tag{4}$$

where $\Phi(z) = \frac{1}{1+exp(-z)}$ is the logistic function, $w_i \in (w_1, \ldots, w_k)$ is the weight and b_D is bias of decision concept D.

To represent the knowledge learned from X_{tr} in the form of QAF, edges between arguments and the concept of decision node D are represented as $E = \{(a_1, D), \ldots, (a_q, D)\}$, and $w_i \in \{w_1, \ldots, w_k\}$ is the strength of edge (a_i, D). Thus, the function ω in QAF_0 can be instantiated as $\omega((a_i, D)) = w_i$, where $w_i \in \{w_1, \ldots, w_k\}$ and $(a_i, D) \in E$. b_D represents the initial score of D. But in a QAF, $\beta(D) \in [0, 1]$, thus we define $\beta(D) = \Phi(b_D)$ according to Eq. (2). In the second step, a MLP model can be described as:

$$s(D) = \Phi\left(\sum_{a_i \in \mathscr{A}_0 \setminus \{a_j, a_k\}} (w_i \times s(a_i)) + b_D + w_c\left(\Phi\left(w_j' \times s(a_j) + w_k{}' \times s(a_k) + b_c\right)\right)\right) \tag{5}$$

where w_i is learned in the last step, thus we fix w_i as a constant score during the parallel training process. w_c is the weight of newly generated concept c_i, and w_j', $w_k{}'$ are new weights of a_j and a_k respectively. b_c is bias of c_i. All the weights and biases can be represented in QAF_i to instantiate ω and β functions.

We continue the parallel learning process until all the strengths of edges and nodes related to newly generated concepts are mined. We obtain a list of fully learned QAF: $(QAF_1, \ldots, QAF_t)$. We evaluate the QAFs in parallel by using X_{vld} and finally select the important concepts by the rule 3. A new description L' for the next P_{QAL} is generated by selecting the descriptions of important concepts stored in L^c and adding the descriptions of ungrouped features or concepts stored in L.

Construct CAM model

Based on the proposed P_{QAL}, we introduce CAM for tabular data decision-making. The construction of CAM is a hierarchical and iterative process, which sequentially stacks Quantitative Argumentation Layers (QALs) to progressively find and select meaningful and important concepts, as illustrated in Fig. 5.

The process begins with the raw feature layer (considered as QAL^0). The first round of P_{QAL} takes the initial features, their descriptions (L), and the data (X) as input. It first generates candidate concepts via the semantic knowledge mining approach and then selects the important ones using the field-wise learning algorithm. This forms the first layer, QAL^1. The output of this layer—comprising the newly formed concepts and any ungrouped features—constitutes a new, more abstract set of nodes with updated descriptions (L^1).

Subsequently, L^1 and the output values from QAL^1 serve as the input for the next round of P_{QAL} to construct a higher-level layer, QAL^2. This stacking process continues, with each new layer building a higher level of semantic abstraction upon the previous one. The construction process of CAM terminates when one of two end conditions is met: (1) No new meaningful concepts can be abstracted during the semantic knowledge mining step. (2) The performance of the base model in the current layer (QAF_0) is lower than that of the previous one, indicating that further abstraction is no longer beneficial.

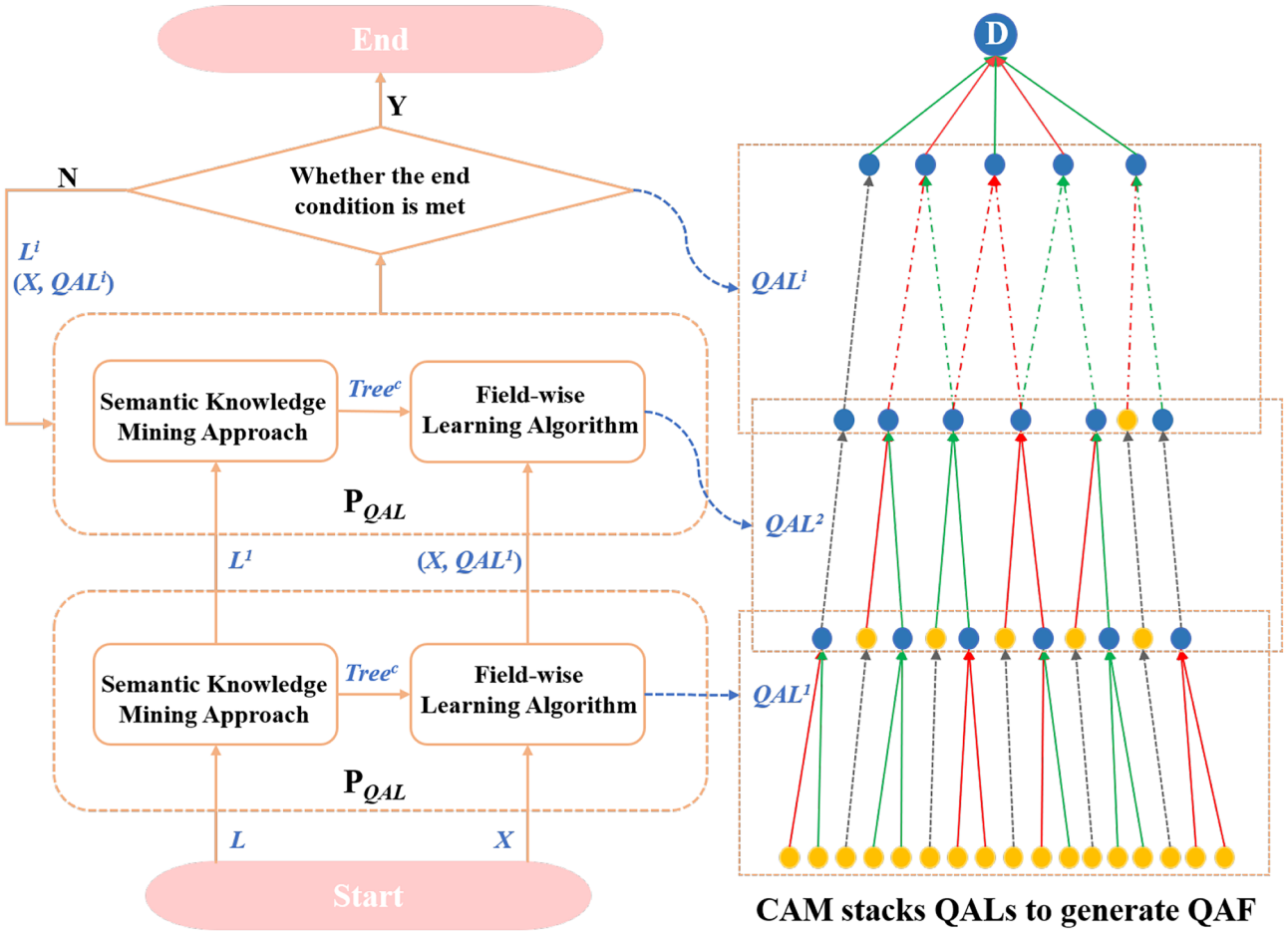


Fig. 5. The detailed workflow of CAM. The left figure illustrates the detailed workflow of CAM, which learns the structure of QALs by P_{QAL}. L^i and QAL^i represents the generated data description and QAL after i-th round P_{QAL}. The right figure shows how the architecture of CAM is built by stacking QALs. The grey dashed lines represent the lower nodes that are not grouped with others and will be omitted in the final CAM structure. The green and red dashed lines represent the omitted stacking processes.

Once the construction is complete, the decision node (*D*) is linked to the final learned *QAL*. The unknown parameters of the entire network (i.e., the weights and biases of the final connections) are then learned, similar to step 1 of the field-wise learning algorithm. This final, integrated hierarchical structure forms the complete quantitative argumentation framework (*QAF*) that constitutes the trained CAM model.

Dialogical explanation within CAM

Dialogical explanation mechanism within CAM

By leveraging the advantages of argumentation structure, CAM is capable of providing the underlying structure for generating dialogical explanations for users. A user may interact with CAM by requesting an explanation of a node (a decision, a concept, or a feature) in CAM. The structure of these explanations, specifically which arguments attack or support a given argument, is defined by the following sets.

Given a tabular data $\mathscr{T} = \langle L, X, D \rangle$, a corresponding $QAF_{\mathscr{T}} : \langle \mathscr{A}_{\mathscr{T}}, E_{\mathscr{T}}, \beta_{\mathscr{T}}, \omega_{\mathscr{T}} \rangle$, an instance $x \in X$ and its decision D with strength $s(D)$, an argumentation dialogue between a user and CAM consists of an explanation request $\mathscr{Q}_{(a)}$ for a node $a \in \mathscr{A}_{\mathscr{T}}$ and an explanation $\mathscr{X}_{(a)}$ which CAM responds with. To generate these explanations, we first define the attacker set $Att(a)$ and supporter set $Sup(a)$ for any argument $a \in \mathscr{A}_{\mathscr{T}}$. The construction of these sets is based on the weights of the edges from its direct child nodes. Specifically, for any direct child node b of a:

- If the edge weight $\omega_{\mathscr{T}}(b, a)$ is positive, b is a supporter of a. The set of all supporters of a is defined as:

$$Sup(a) = \{b \in \mathscr{A}_{\mathscr{T}} \mid (b, a) \in E_{\mathscr{T}} \wedge \omega_{\mathscr{T}}(b, a) > 0\} \tag{6}$$

- If the edge weight $\omega_{\mathscr{T}}(b, a)$ is negative, b is an attacker of a. The set of all attackers of a is defined as:

$$Att(a) = \{b \in \mathscr{A}_{\mathscr{T}} \mid (b, a) \in E_{\mathscr{T}} \wedge \omega_{\mathscr{T}}(b, a) < 0\} \tag{7}$$

The user's explanation request $\mathscr{Q}_{(a)}$ is answered by the explanation $\mathscr{X}_{(a)}$, which is constructed based on the strength of the arguments within these defined sets. According to Eqs. (1) and (2), for any argument $a \in \mathscr{A}$, its strength $s(a)$ can be obtained by following formula:

$$s(a) = \Phi\left(ln\left(\frac{\beta_{\mathscr{T}}(a)}{1 - \beta_{\mathscr{T}}(a)}\right) + \sum_{b \in Att(a)} \omega_{\mathscr{T}}(b, a) \times s(b) + \sum_{b \in Sup(a)} \omega_{\mathscr{T}}(b, a) \times s(b)\right) \tag{8}$$

Our intuition is that the dialogical explanation is simpler than but consistent with CAM. The explanation of an argument a may consist of its supporters and attackers, which have significant impacts on a. Therefore, we propose a mechanism to simplify the structure of CAM, called the SSC algorithm by searching the dominant arguments, described in Algorithm 2. For any $S \subseteq \mathscr{A}$, if $S = \emptyset$, let $max(S) = \emptyset$; else, let $max(S) = argmax_{b \in S}(|\omega(b, a) \times s(b)|)$, where *argmax* refers to the argument b, at which the absolute value of $(\omega_{\mathscr{T}}(b, a) \times s(b))$ is as large as possible. In Algorithm 2, we aim to find the least supporters and attackers in $Sup(a)$ and $Att(a)$ with significant impacts on a to make the same decision with CAM, and the sets containing the selected supporters and attackers are denoted as $Sup'(a)$ and $Att'(a)$, respectively.

Input: a, $s(a)$, $Att(a)$, $Sup(a)$
Output: $Att'(a)$, $Sup'(a)$
if $Att(max(Att(a))) \cup Sup(max(Att(a))) = \emptyset$ **then**
  $Att'(a) \Leftarrow max(Att(a))$;
else
  $b = max(Att(a))$;
  SSC(b, $s(b)$, $Att(b)$, $Sup(b)$);
end
if $Att(max(Sup(a))) \cup Sup(max(Sup(a))) = \emptyset$ **then**
  $Sup'(a) \Leftarrow max(Sup(a))$;
else
  $b = max(Sup(a))$;
  SSC(b, $s(b)$, $Att(b)$, $Sup(b)$);
end
$s'(a) = \Phi\left(\ln\left(\frac{\beta(a)}{1-\beta(a)}\right) + \sum_{b \in Att'(a)} \omega(b,a)s(b) + \sum_{b \in Sup'(a)} \omega(b,a)s(b)\right)$;
if $s(a) > 0.5$ *and* $s'(a) \leq 0.5$ **then**
  while $s'(a) \leq 0.5$ **do**
    $c = \max\{b \mid b \in Sup(a) \wedge b \notin Sup'(a)\}$;
    if $Att(c) \cup Sup(c) = \emptyset$ **then** $Sup'(a) \Leftarrow c$;
    else SSC(c, $s(c)$, $Att(c)$, $Sup(c)$);
    update $s'(a)$;
  end
  return $Att'(a)$, $Sup'(a)$.
else if $s(a) \leq 0.5$ *and* $s'(a) > 0.5$ **then**
  while $s'(a) > 0.5$ **do**
    $c = \max\{b \mid b \in Att(a) \wedge b \notin Att'(a)\}$;
    if $Att(c) \cup Sup(c) = \emptyset$ **then** $Att'(a) \Leftarrow c$;
    else SSC(c, $s(c)$, $Att(c)$, $Sup(c)$);
    update $s'(a)$;
  end
  return $Att'(a)$, $Sup'(a)$.
else
  return $Att'(a)$, $Sup'(a)$.

Algorithm 2. SSC algorithm

Then, according to the simplified argumentation structure, we provide a simple argumentation dialogue for risk assessment as follows. Building on prior work in argumentation theory[16], we define two phrase-generating functions, $r^{safe}(a)$ and $r^{risky}(a)$, for any argument $a \in \mathscr{A}$. These functions are based on l_a, which is a natural language description of the feature or concept that argument a represents. This natural language representation makes the model's internal reasoning comprehensible to a human user. The determination of whether an argument's value is "good" or "bad" is based on its impact on the target outcome's risk. Specifically, if an argument's value increases the risk of the target outcome, it is considered "bad" and triggers the "risky" response template. Conversely, if it decreases the risk, it is considered "good" and triggers the "safe" response template.

$$r^{safe}(a) = (\text{the})\ l_a\ \text{was good};$$
$$r^{risky}(a) = (\text{the})\ l_a\ \text{was bad};$$
$$r^{risky}(\varnothing) = r^{safe}(\varnothing) = \{\}.$$

We acknowledge that this dialogue template is specifically designed for the risk assessment problem to align with the application context and model logic. For different application domains, a new set of corresponding templates would be designed. Since this paper focuses on the risk assessment problem, we provide this specific template as a concrete example, demonstrating how our framework can be applied to a real-world task. An *argumentation dialogue* is such that for any $a \in \mathscr{A}$:

$$
\begin{aligned}
&\text{if } a = D \text{ and } s(D) > 0.5\text{:}\\
&\quad \mathscr{Q}(a) = \{\text{Why was } a \text{ assessed as high risk?}\}\\
&\quad \mathscr{X}(a) = \{\text{This case was assessed as high risk because}\}\\
&\qquad\qquad + \sum_{b \in Sup'(a)} r^{risky}(b) + \{although\} + \sum_{b \in Att'(a)} r^{safe}(b);\ \text{else}\\
&\text{if } a = D \text{ and } s(D) \leq 0.5\text{:}\\
&\quad \mathscr{Q}(a) = \{\text{Why was } a \text{ assessed as low risk?}\}\\
&\quad \mathscr{X}(a) = \{\text{This case was assessed as low risk because}\}\\
&\qquad\qquad + \sum_{b \in Att'(a)} r^{safe}(b) + \{although\} + \sum_{b \in Sup'(a)} r^{risky}(b);\ \text{else}\\
&\text{if } a \in \mathscr{C} \text{ and } a \in Sup(D)\text{:}\\
&\quad \mathscr{Q}(a) = \{\text{Why was } a \text{ considered to be bad?}\}\\
&\quad \mathscr{X}(a) = \{l_a \text{ was considered to be bad because}\}\\
&\qquad\qquad + \sum_{b \in Sup'(a)} r^{risky}(b) + \{although\} + \sum_{b \in Att'(a)} r^{safe}(b);\ \text{else}\\
&\text{if } a \in \mathscr{C} \text{ and } a \in Att(D)\text{:}\\
&\quad \mathscr{Q}(a) = \{\text{Why was } a \text{ considered to be good?}\}\\
&\quad \mathscr{X}(a) = \{l_a \text{ was considered to be good because}\}\\
&\qquad\qquad + \sum_{b \in Att'(a)} r^{safe}(b) + \{although\} + \sum_{b \in Sup'(a)} r^{risky}(b);\ \text{else}\\
&\text{if } \ a \in \mathscr{F}\text{:}\\
&\quad \mathscr{Q}(a) = \{\text{Why was } a \text{ considered to be bad (or good)?}\}\\
&\quad \mathscr{X}(a) = \{\text{Because in this case, } l_a \text{ was } x_a\};
\end{aligned}
$$

where *Att*(*D*) and *Sup*(*D*) represent arguments attacking or supporting *D* in a broad sense, and these arguments are not necessarily directly connected to *D*, x_a denotes the input value of feature *a*.

Case study: explanation for risk estimation

Consider CAM in Fig. 6 for a risk assessment case (which is labeled as high risk by the bank) from the Fico dataset.[2] The QAF in the figure is built up by extracting the concepts and dialectical relations between them from raw features with the help of P_{QAL}. The key paths for reasoning searched by the SSC algorithm are thickened, starting with the strongest supporting and attacking arguments "Inquiry" and "ExternalRiskEstimation".

The base score of the decision node "Risk" is initially set to 0.5, while the base scores of other concept nodes are calculated by field-wise algorithm in CAM. According to the MLP-based reasoning method, the strength for "Risk" with original QAF is $s(Risk) = 0.84$, which is seen as a "high-risk case" because $s(Risk) > 0.5$. With the help of the SSC algorithm, the complex QAF can be simplified as shown in Fig. 6b. The strength for "Risk" with simplified QAF is $s(Risk) = 0.54$, which has the same decision as the original one but with a more intuitive and simpler reasoning process. A simple argumentation dialogue between a user and CAM may then be as follows:

User: Why was this case assessed as high risk?

CAM: This case was assessed as high risk because the information about this consumer's credit bureau report pulled by a lending institution (the description of "Inquiry") was bad, although the consolidated safe score (the description of "ExternalRiskEstimation") was good.

User: Why was the "Inquiry" considered to be bad?

CAM: The "Inquiry" was considered to be bad because the number of times that a lending institution has pulled this consumer's credit bureau report (the description of "NumInquiry") was bad, although the month since the most recent inquiry excluding 7 days (the description of "MsinceMostRecentInqexcel7days") was good.

User: Why was the "NumInquiry" considered to be bad?

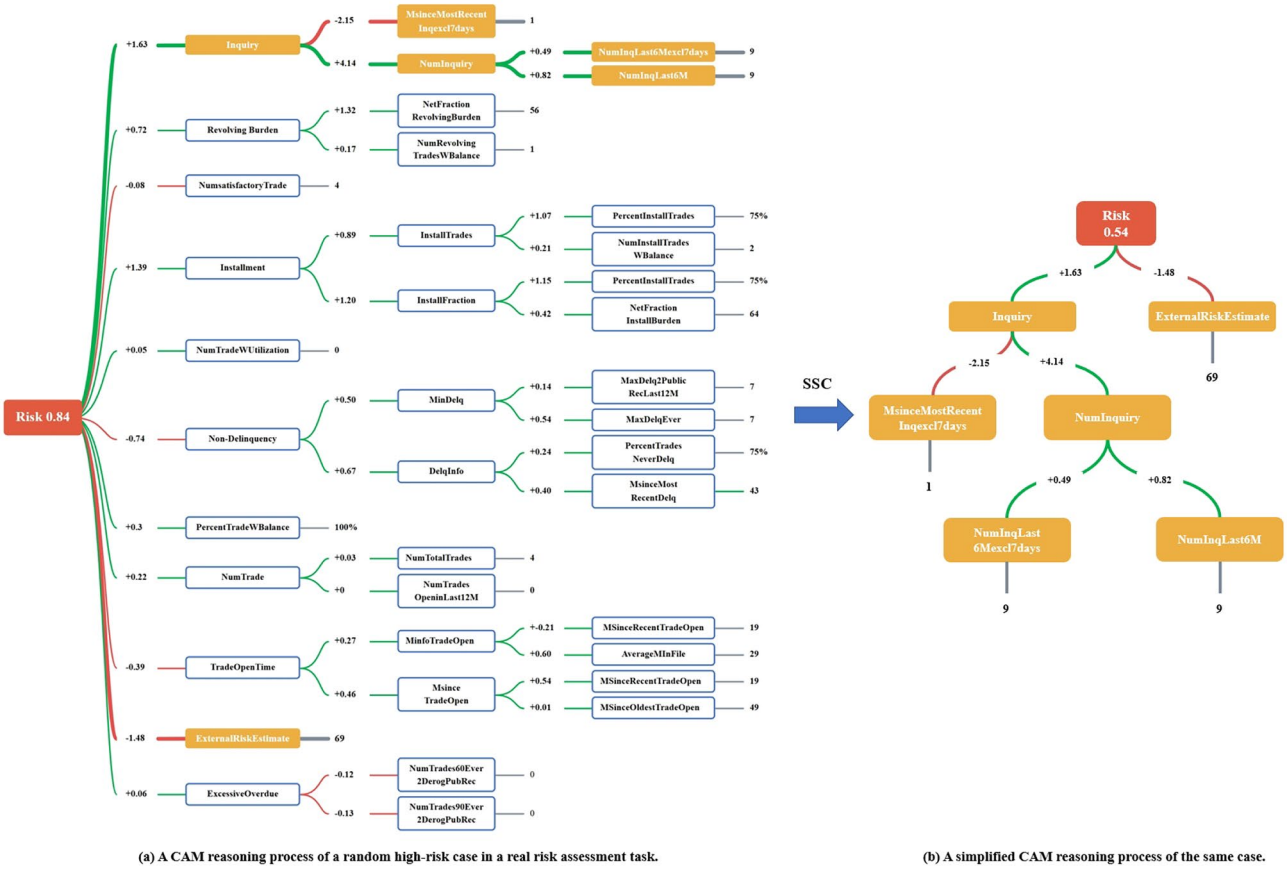
CAM: The "NumInquiry" was considered to be bad because the number of inquiries in the last 6 months (the description of "NumInqLast6M") was bad, and the number of inquiries in the last 6 months excluding 7 days (the description of "NumInqLast6Mexcel7days") was bad.

User: Why was the "NumInqLast6M" considered to be bad?

CAM: Because the number of inquiries in the last 6 months was 9.

Analysis From the explanation, we know that "Inquiry" has the largest positive influence on "Risk", while "ExternalRiskEstimation" has the largest negative influence. From the perspective of banks, the number of inquiries increased because customer actively applied for a new credit card or mortgage. Researches show that opening several credit accounts in a short period represents greater credit risk. In this case, this customer applied 9 times for new credit cards or mortgages within 6 months, even though in the last month there was

[2] https://community.fico.com/s/explainable-machine-learning-challenge/.



(a) A CAM reasoning process of a random high-risk case in a real risk assessment task.

(b) A simplified CAM reasoning process of the same case.

Fig. 6. A visualization of a random high-risk case in a risk assessment task. (**a**) illustrates that CAM predicts a risk score of 0.84 for this case. (**b**) shows that the predicted risk score is 0.54 by using the simplified CAM. The feature values of this case are input through the grey lines. The green and red edges denote the support and attack relations, respectively. The value labeled on each edge represents the influence from the lower node to the upper node, calculated as the product of the lower node value and the edge value.

no new application record, the "Inquiry" information of this customer still looks bad. Although the bank has a higher evaluation score (69) for this user, the influence of "Inquiry" is bigger than "ExternalRiskEstimation", the customer is still assessed as a high risk for delinquency. The explanation shows the key path of the reasoning process, it is more intuitive and straightforward. Moreover, the human knowledge within the explanation makes it more acceptable and understandable.

Experiments

User study

The setting of user study

We conducted a user study with 134 banking experts on an online platform ("Survey Star"[3] , Changsha Ran Xing Science and Technology, Shanghai, China) to evaluate three aspects: (1) the comprehensibility of CAM, (2) the alignment of CAM explanations with human decision logic, and (3) the preference for CAM compared to other explanation methods based on feature importance. The study focused on the Fico dataset (see Appendix). Banking experts were chosen because CAM requires a strong understanding of the domain knowledge. Eligibility criteria included: (1) a minimum of a Bachelor's degree, (2) a finance-related major, and (3) a bank-related job. Figure 7 summarizes the backgrounds of the 134 participants. Participants were randomly assigned to evaluate either CAM (n = 66) or EBM (n = 68), with each participant reviewing explanations for randomly selected samples. Using separate participant groups ensured an independent assessment of each explanation model and avoided potential bias. EBM was selected as the comparison method due to its state-of-the-art explainability and ability to provide local explanations. We calculated the mean and standard deviation of all metrics and assessed statistical significance using t-tests. This approach allowed us to rigorously compare the performance of CAM and EBM.

[3] https://www.wjx.cn.

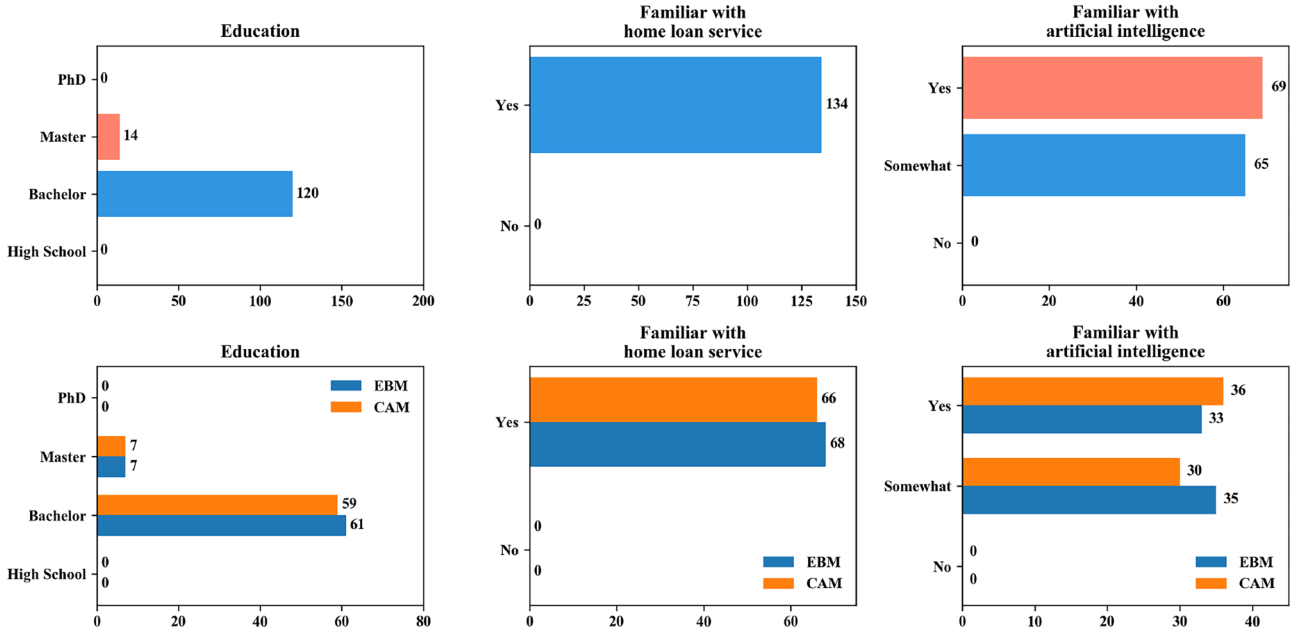


Fig. 7. Backgrounds of participants.

Aspect	Question/task	Result
Global comprehension	Selection of reasonable feature combinations	85% of participants agreed with CAM
	Prediction of argumentative relations (attack/support)	91% (among participants who picked the same features as CAM) matched CAM
Local comprehension	Decision-making using input data only	39% accuracy
	Decision-making using CAM visualization without outputs	70% accuracy
	Decision-making using CAM visualization with outputs	70% accuracy
Perceived helpfulness	Can argumentative relations improve understanding?	21% "Somewhat helpful", 50% "Helpful", 29% "Very helpful" (Avg. = 4.07/5)

Table 1. User study results on comprehensibility of CAM.

All methods were carried out in accordance with relevant guidelines and regulations. The Ethics Review Board in the Center for Psychological Sciences at Zhejiang University, China, approved our study. Informed consent was obtained electronically from all participants prior to their involvement in the study.

Comprehensibility

To evaluate whether humans can understand both the global structure and local explanations provided by CAM, we conducted a user study focusing on three aspects: global comprehension, local comprehension, and perceived helpfulness (Table 1).

Global comprehension Participants were asked to select reasonable feature combinations and predict the argumentative relations (attack or support) among CAM's components. *Results*: 85% of participants selected the same feature combinations as CAM. Among those, 91% correctly predicted the argumentative relations in agreement with CAM. These findings indicate that users can effectively understand the overall structure of CAM, including both nodes and their relations.

Local comprehension To evaluate participants' ability to interpret individual decisions, we compared three settings: input data alone, CAM visualizations without outputs, and CAM visualizations with outputs (e.g., Fig. 6a). *Results*: With input data alone, only 39% of participants reached correct decisions. This accuracy, being notably below the random-guess baseline of 50% for this near-balanced dataset, suggests that the risk assessment task is non-trivial for humans when based solely on raw feature values. The complex, non-linear interactions among the 23 features likely pose significant challenges for manual cognitive integration, underscoring the need for model-based decision support. Accuracy increased to 70% when CAM visualizations without outputs were provided, while adding outputs did not yield further improvement. This indicates that the argumentative structure conveyed by CAM already plays the dominant role in supporting local interpretability, whereas the additional outputs contributed little beyond this effect.

Aspect	Category	CAM	EBM	*p* value
Human precision	Reasonableness	4.09 ± 0.63	2.65 ± 1.38	$p < 0.001$
	Overlap of important features	0.63 ± 0.13	0.44 ± 0.11	$p < 0.001$
User acceptance	Ease of understanding	3.86 ± 0.68	3.83 ± 0.61	0.82
	More helpful knowledge	4.35 ± 0.64	3.90 ± 0.67	$p < 0.001$
	Increases confidence	4.10 ± 0.69	3.68 ± 0.67	$p < 0.001$

Table 2. Comparison of CAM and EBM on human precision and user acceptance (mean ± std, t-test).

Model	Stability	Stability_CI	Fidelity	Fidelity_CI	Faithfulness	Faithfulness_CI
EBM	0.984	(0.789, 1.000)	0.876	(0.000, 1.000)	0.876	(0.000, 1.000)
CAM	0.957	(0.397, 1.000)	1.000	(1.000, 1.000)	1.000	(1.000, 1.000)

Table 3. Comparison of mean explanation metrics across 5 random seeds (mean top-k for CAM: 2.63).

Perceived helpfulness Participants were asked to rate how much the argumentative relations of the local explanation helped them understand the decisions. *Results*: 21% rated the explanations as Somewhat helpful", 50% as Helpful", and 29% as "Very helpful", yielding an average score of 4.07 out of 5. These results confirm that users can leverage the argumentative structure of CAM to support decision-making effectively.

Human precision

Human precision measures the extent to which model explanations align with human decision logic[26] (Table 2) We evaluated this aspect at two levels: *Reasonableness of explanations.* Participants reviewed whether the explanations provided reasonable rationales and rated them on a five-point Likert scale (1 = "not reasonable at all", 5 = "very reasonable"). *Result:* CAM achieved significantly higher scores than EBM (4.09 ± 0.63 vs. 2.65 ± 1.38, $p < 0.001$), indicating that CAM offers explanations that are more consistent with human judgment. *Overlap of important features.* At the feature level, the best explanations highlight the most relevant and least irrelevant features[26]. Participants selected multiple features they considered most important, and only selections from participants who made correct decisions were analyzed. We then identified the top four features most frequently selected by participants and computed their overlap with the top four features identified by each model. *Result:* CAM exhibited a higher overlap with participants' selections (0.63 ± 0.13) than EBM (0.44 ± 0.11, $p < 0.001$), demonstrating that CAM's explanations better align with the features humans consider critical. CAM's explanations are valuable because they not only highlight these critical features but also transparently show their quantitative contributions and dialectical relationships within the reasoning hierarchy.

User acceptance

User acceptance evaluates how understandable, informative, and confidence-enhancing the model explanations are for human users (Table 2). *Ease of understanding.* Participants rated how easily they could comprehend the explanations on a five-point Likert scale. *Result:* CAM (3.86 ± 0.68) and EBM (3.83 ± 0.61) showed no significant difference ($p = 0.82$). We posit that this is because CAM provides more information that takes more time to process compared to the EBM model. The additional useful knowledge leads to a decrease in the ease of understanding of CAM. Even so, CAM still scored higher on ease of understanding compared to EBM. We think that this is because the feature combinations without any semantics in EBM may lead to confusion for the users. *More helpful knowledge.* Participants evaluated whether the explanations provided additional useful knowledge. *Result:* CAM received higher scores than EBM (4.35 ± 0.64 vs. 3.90 ± 0.67, $p < 0.001$), suggesting that CAM conveys more actionable and informative insights. *Increases confidence.* Participants rated whether the explanations increased their confidence in decision-making. *Result:* CAM again outperformed EBM (4.10 ± 0.69 vs. 3.68 ± 0.67, $p < 0.001$), indicating that CAM explanations are more effective in supporting human trust and confidence.

Objective explanation metrics

To better evaluate the explanations, we conducted additional experiments using five random seeds on the Fico dataset. For each seed, we computed Stability, Fidelity, and Faithfulness for both CAM and EBM. These metrics are accompanied by their 95% confidence intervals (CI). *Stability* measures the robustness of the explanation under small input perturbations. For both CAM and EBM, Gaussian noise with a coefficient of 1×10^{-3} was added to each feature, and the fraction of unchanged top-k features was recorded. *Fidelity* quantifies how well the explanation reflects the model's decision, computed as the fraction of predictions preserved when only the top-k features are retained. *Faithfulness* evaluates whether the top-k features sufficiently influence the model output, measured as the change in predictions when the top-k features are masked. Top-k features refer to the number of feature nodes in the minimal subset computed by Algorithm 2, which are sufficient to explain the model's prediction.

Table 3 shows that Algorithm 2 in CAM identifies a minimal feature subset for each prediction. Consequently, Fidelity and Faithfulness are consistently perfect (1.0), indicating that explanations are highly faithful and

Source type	Name	Domain	#Samples	#Features	Positive rate (%)
Benchmarks	Fico	Banking	9871	23	52.03
	Mimic3	Healthcare	27,348	57	9.83
In domain dataset	data1	E-commence	96452	33	3.2
	data2	E-commence	98,936	65	2.0

Table 4. Summary of datasets.

Model	Hyperparameter settings
CAM	Target encode categorical features; Quantile transform all features; Text preprocessing: remove special characters, embed with multilingual Sentence-BERT; Cosine similarity for clustering with threshold $\mu = 0.55$; Augmentation-based field-wise filter fine-tuned for 5 epochs
Logistic Regression (LR)	LBFGS optimizer; single large batch; L2 regularization with default weight
XGBoost (XGB)	Hyperparameters adopted from Node-GAM[7] to ensure full convergence
EBM	Default parameters; maximum rounds = 20,000
Artificial Neural Network (ANN)	3-layer MLP (128, 64 hidden units); Batch normalization; LeakyReLU activation
NODE-GA2M	Second-order interaction mode; default values for other hyperparameters
NBM	3-layer MLP (256, 128, 128 hidden units); ReLU activation; 100 basis outputs
GRAND-SLAMIN	Default hyperparameters as specified in the original paper[9]

Table 5. Hyperparameter settings for compared models.

complete, even with fewer than three features (2.63). In contrast, the same number of top-k features in EBM does not achieve comparable reliability. CAM's Stability is slightly lower (0.957 vs. 0.984 for EBM), reflecting some sensitivity to input perturbations. This difference is expected because EBM uses feature binning internally, which enhances robustness to small noise and leads to better stability in the perturbation experiments. This observation also provides insights for potential future improvements in CAM's stability mechanism.

Notably, CAM has been applied to Alibaba's e-commerce applications to explain potential fraud cases. The feedback shows that auditors of Alibaba prefer explanations based on human-level knowledge rather than explanations based on feature importance.

Data experiments

Introduction of dataset

We evaluate CAM's performance on four datasets, comprising two public benchmarks and two proprietary industrial datasets. The first public benchmark is the *Fico* dataset, which originates from the banking domain. The predictive task is to perform credit risk assessment based on a consumer's credit bureau data. The second is the *Mimic3* clinical dataset. For this dataset, the task is to predict in-hospital mortality using physiological data collected within the first 24 h of a patient's stay in the intensive care unit (ICU). In addition to the public benchmarks, we use two in-domain, anti-fraud datasets, denoted as *data1* and *data2*, collected from two Alibaba e-commerce applications. All datasets are medium-sized, with 10,000–100,000 samples. Table 4 summarizes the characteristics of each dataset, and further details are provided in the Appendix.

Compared experiments

The setting of compared experiment We split the dataset into training, validation, and testing sets with a 64–16–20 ratio and repeated each experiment with five random seeds. All tasks are binary classification, and the evaluation metric is AUC. The compared models are categorized into three types: *Transparent models* (CAM, Logistic Regression) require the model's reasoning process to be fully observable, allowing users to make decisions based on model information. *Interpretable models* (EBM, NODE-GAM, NBM, GRAND-SLAMIN) provide appropriate explanations of the decision-making process but do not require full transparency. *Black-box models* (XGBoost[27], Artificial Neural Network) require neither transparency nor explanations usable by humans for decision-making. These models were chosen because they are commonly used for tabular data classification. Among them, XGBoost is recognized for strong predictive performance, and EBM, NODE-GA2M, NBM, and GRAND-SLAMIN are state-of-the-art interpretable models. The hyperparameter selections are shown in Table 5.

Analysis on compared results In Table 6, we report the comparative results of CAM, LR, EBM, NODE-GA2M, NBM, GRAND-SLAMIN, ANN, and XGB in terms of the mean and standard deviation (std) of AUC.[4] CAM achieves the highest mean AUC on the Fico dataset, GRAND-SLAMIN on Mimic3, and XGB on the other two datasets. On average, CAM ranks fourth, close to EBM and behind XGB and GRAND-SLAMIN. Regarding

[4]Mean is the average of the five experimental runs, reflecting overall performance, while std indicates stability.

	Transparent models		Interpretable models				Black-box models	
	CAM	LR	EBM	NODE-GA2M	NBM	GRAND-SLAMIN	ANN	XGB
Fico	**80.23** ± 0.92	79.74 ± 1.14	80.09 ± 0.87	77.22 ± 0.81	76.30 ± 0.83	80.07 ± 1.05	77.22 ± 0.81	79.88 ± **0.78**
Mimic3	79.76 ± 0.49	77.66 ± 0.80	80.71 ± 1.18	81.22 ± 0.64	61.66 ± **0.12**	**82.11** ± 0.76	72.71 ± 2.16	81.99 ± 1.21
data1	93.12 ± 0.23	86.04 ± 0.49	94.96 ± **0.12**	92.78 ± 0.66	94.63 ± 0.17	94.78 ± 0.26	82.61 ± 1.85	**97.13** ± 0.40
data2	96.16 ± 0.24	83.94 ± 4.15	96.17 ± 0.20	96.67 ± **0.14**	89.08 ± 0.87	96.30 ± 0.15	71.54 ± 12.78	**97.14** ± 0.26
Average	87.32 ± **0.47**	81.85 ± 1.64	87.98 ± 0.59	86.97 ± 0.56	80.42 ± 0.50	88.32 ± 0.56	76.02 ± 4.40	**89.04** ± 0.66

Table 6. Mean and standard deviation (std) of AUC (%) results for four datasets in 5 random seeds experiments conducted by CAM and other machine learning approaches. The best results are highlighted in bold.

Model	Mean AUC (%)	Std AUC (%)
CAM with correlation-based combinations	80.16	1.07
CAM with logistic regression algorithm	80.22	0.98
CAM	**80.23**	**0.92**

Table 7. Ablation results for the Fico datasets in 5 random seeds experiments. The best results are highlighted in bold.

stability (std), XGB performs best on Fico, NBM on Mimic3, EBM on data1, and NODE-GA2M on data2, while CAM achieves the best overall stability. We attribute this to its structure grounded in human-level knowledge. Overall, CAM demonstrates competitive accuracy and high stability compared with transparent, interpretable, and black-box models.

Ablation experiment

Semantic knowledge mining approach We evaluate CAM on the Fico dataset using two feature combination strategies: semantic knowledge mining and correlation-based grouping (Table 7). CAM with semantic mining achieves higher mean AUC and notably lower variance, indicating that human knowledge leads to more stable and robust feature abstractions than purely data-driven methods.

Beyond this controlled setting, we also test CAM in real-world e-commerce data. Even when feature descriptions are sparse or absent, CAM can utilize semantic cues from feature names. This is a crucial aspect for practical applications, as high-quality, detailed feature descriptions are not always available. By falling back on feature names (e.g., 'NumInqLast6M', 'ExternalRiskEstimate'), which often contain strong semantic signals, CAM can still construct a meaningful initial concept structure. While the quality of descriptions directly impacts the quality of the initial semantic grouping, this flexibility allows CAM to be applied even in scenarios with imperfect documentation. In practice with Alibaba Group, this proved effective for fraud risk prediction, demonstrating that meaningful explanations can still be obtained without detailed descriptions.

Field-wise learning algorithm In Table 7, we compare CAM predictions on the Fico dataset with different learning approaches, i.e., field-wise learning algorithm and logistic regression algorithm. On the mean AUC, these two learning approaches show similar performances. The results indicate that the field-wise learning strategy did not compromise the performance of CAM. CAM with the field-wise learning algorithm has a small std improvement over the logistic regression algorithm. We assume that rule 3 in the field-wise learning algorithm makes the CAM structure more stable.

The ablation study (Table 7) further confirms that the performance gain is not solely from clustering semantics but from the interaction between semantic abstraction and the quantitative argumentation process. When clustering is replaced by correlation-based grouping, performance and stability drop, suggesting that the integration of textual semantics into the QAF reasoning structure–rather than clustering itself–drives the improvement.

Efficiency experiments

The setting of efficiency experiment In order to measure the parallel efficiency of the field-wise learning algorithm in CAM model as well as to verify whether CAM model is capable of performing modeling work on large datasets, we designed a set of experiments in which the parallel lines of CAM model were set to 1, 2, 4, 8, and 16 on datasets with sizes of 1000, 10,000, 100,000, and 1,000,000.

Analysis on runtime results The runtime results, as shown in Table 8 and Fig. 8, demonstrate that the algorithm scales effectively, particularly for large datasets. While speedup increases with the core count, it eventually saturates, with a corresponding decrease in parallel efficiency due to communication and synchronization overheads. Despite this, the runtime on a large-scale dataset of 1,000,000 instances significantly improves. It decreases from over 4 h (14,726 s) with a single core to a more manageable 38 min (2306 s) with 16 parallel cores.

Problem size	Cores	Duration (s)	Speedup	Efficiency
1000	1	11.59	1.00	1.00
	2	7.28	1.59	0.80
	4	5.09	2.27	0.57
	8	4.20	2.76	0.35
	16	4.16	2.78	0.17
10,000	1	110.66	1.00	1.00
	2	59.12	1.87	0.94
	4	34.29	3.23	0.81
	8	22.15	4.99	0.62
	16	20.62	5.37	0.34
100,000	1	1466.85	1.00	1.00
	2	775.34	1.89	0.95
	4	428.67	3.42	0.85
	8	257.30	5.70	0.71
	16	227.55	6.45	0.40
1,000,000	1	14726.61	1.00	1.00
	2	7793.73	1.89	0.95
	4	4343.37	3.39	0.85
	8	2652.39	5.55	0.69
	16	2306.92	6.38	0.40

Table 8. Performance Metrics for Various Problem Sizes.

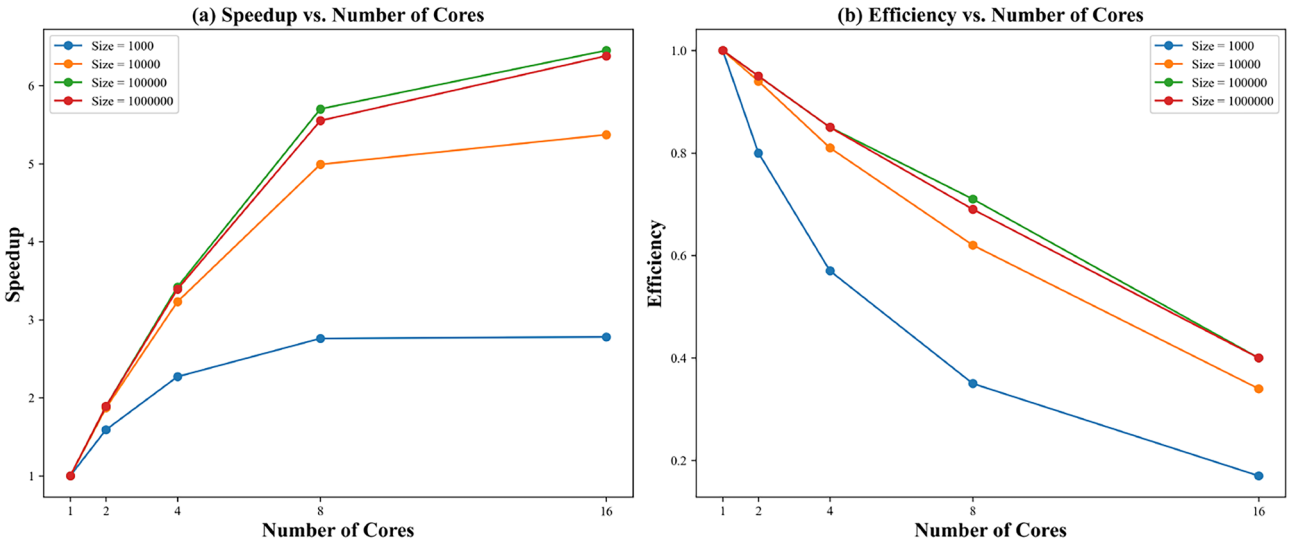


Fig. 8. Speedup and Efficiency of CAM with various magnitudes of data and multiple numbers of parallel cores.

This substantial reduction in modeling time proves that the algorithm is capable of handling large-scale data in a practical timeframe.

Experiments of imbalance data

We evaluate the robustness of CAM on different positive rates on four datasets. The positive rates of datasets are set as 0.0625, 0.125, 0.25, 0.5, and 1. We repeat the experiments with five random seeds, and the AUC results in Fig. 9 are the mean value of the 5 experiments. On the Fico and data2 datasets, CAM performances barely fluctuate, while on the Mimic 3 and data1 datasets, the differences in CAM performances are less than 0.002. The results indicate that CAM is not sensitive to imbalanced data.

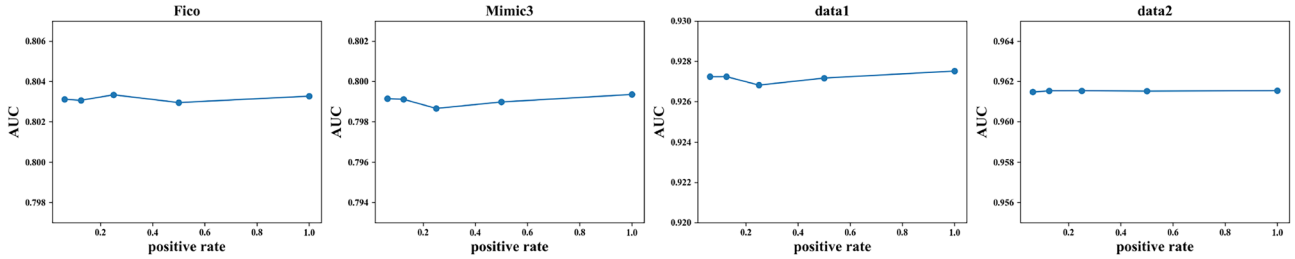


Fig. 9. Performance of CAM with different data imbalance on four datasets.

Related work

Concept-based models

Concept-based models (CBMs) map the black-box visual representations extracted by deep neural networks onto a set of interpretable concepts and use the concepts to make predictions, enhancing the transparency of the decision-making process. Early research has focused on generating high-level human concepts from image data for explainable prediction, such as TCAV[28], VCEC[29], ProtoPNet[30], and ACE[25]. Research on CBMs has been maturing rapidly[31–35]. Furthermore, CBMs have expanded into the realm of natural language processing, where many language models[36–38] introduced human-interpretable concepts into deep learning networks to enhance model interpretability. Recent advancements like Probabilistic Concept Bottleneck Models[34] have further refined CBMs for handling uncertainty. However, these methods, like their predecessors, are predominantly designed for unstructured data. Our work addresses this gap by proposing a novel framework to automatically construct and leverage a concept hierarchy specifically for tabular data.

Feature-based XAI research for tabular data

For tabular data, a group of post-hoc methods explain models by computing or approximating the feature importance, such as LIME and SHAP. However, these explanations are known to be unstable and unfaithful[5,39]. Recently, researchers have focused on constructing an interpretable model. NODE-GAM, NBM, GRAND-SLAMIN, and EBM belong to a class of interpretable models that can provide feature-based explanations. These methods apply neural network models and boosting tree models to fit a functional relationship between features and the output. However, it has been argued that the same data can yield divergent or conflicting interpretations depending on the fitting model used[40]. This is because the fitting model relies solely on a data-driven mechanism and may lead to overfitting. DANETs[41] can abstract higher-level tabular features by neural networks, but the higher-level features only contribute to the classification performance. The semantics within them are not completely explicit, which may lead to confusing features.

Counterfactual explanations for tabular data

Counterfactual explanations (CFEs) are widely used as a post-hoc, local explanation technique for tabular data. Unlike feature-importance methods such as SHAP or LIME, CFEs adopt a user-centered and intuitive "what–if" approach. Many studies have focused on identifying the closest counterfactual instances in the feature space[42] or on improving the efficiency of this search process[43]. However, research[44] has shown that different CFE algorithms may yield conflicting explanations for the same black-box model, raising serious ethical and practical concerns. Furthermore, Guidotti[45] reviewed and benchmarked CFE methods, arguing that relying solely on optimization to generate counterfactuals does not necessarily improve user understanding and may even be misleading. As a result, recent work[46] has emphasized shifting the focus of CFEs from computational efficiency and objective metrics toward human-centered criteria such as user acceptance and comprehensibility. It is important to note that CAM provides a different paradigm of explanation. Instead of generating 'what–if' scenarios like CFEs, CAM offers a structural, dialogical explanation that reveals the model's internal reasoning hierarchy, addressing the user's need for 'how' the decision was made rather than 'what' needs to change.

Argumentative XAI

There are two types of argumentative XAI. One is post-hoc argumentative explanations, which apply argumentation frameworks (AFs) to models that are not inherently argumentative. For example, some studies construct AFs using rules extracted from neural networks[47], while others represent neural networks as quantitative bipolar argumentation frameworks (QBAFs) by interpreting groups of neurons as arguments[48]. Additionally, argumentative explanations for machine learning classifiers have been proposed using argumentation-based explanation functions[13,49]. And some researchers offer meta-explanations where AFs justify the explanations themselves[50]. A primary drawback of these approaches is that they do not fully and accurately reflect the internal workflow of the underlying models.

The other comprises argumentation-based interpretable models, where the models themselves are built as AFs. ArgEML is a representative framework that couples ML-based rule initialization with reasoning executed in SWI-Prolog[51]. This ensures logical consistency and clear explanations, but scales poorly as the rule base grows. In contrast, decision-making models[14–17] combine machine learning with argumentation semantics to improve scalability and enable argumentative reasoning in data-intensive applications. However, these models inherently rely on a pre-existing knowledge structure. To the best of our knowledge, no existing work has introduced a

method that can automatically construct an argumentative structure from tabular data, which is essential for applying this approach universally without relying on domain experts or specific data formats.

Hierarchical and structured explanations

Our work is related to the broader literature on interpretable models with structured explanations. For instance, MUSE[52] explains black-box models by generating a set of "if–then" rules organized as a decision set, which provides faithful and customizable explanations. However, CAM differs from MUSE and similar post-hoc methods in several fundamental ways. First, CAM is an intrinsically interpretable model (a white-box), not a post-hoc explainer for an existing black-box model. Its structure is the model itself. Second, CAM's construction is knowledge-guided, leveraging textual descriptions to form its semantic hierarchy, whereas MUSE is purely data-driven. Third, the underlying explanatory paradigm is different. CAM is built on a Quantitative Argumentation Framework, which naturally represents dialectical relationships (support vs. attack) and enables interactive, dialogical explanations. This contrasts with the decision-rule format of MUSE. Our contribution lies in creating a model that automatically builds this argumentative structure from tabular data, a previously unaddressed challenge.

Conclusions

In this work, we present CAM as a human-aligned interpretable framework for tabular data in high-risk domains. CAM integrates semantic abstraction from data descriptions with quantitative argumentation reasoning to construct interpretable reasoning structures. Through this integration, descriptive textual knowledge is transformed into human-understandable concepts that participate in explicit support–attack relations within a quantitative argumentation framework.

The dialogical explanations derived from these structures allow CAM to provide transparent, stepwise reasoning paths consistent with human decision logic. Human evaluations indicate that CAM achieves a high level of comprehensibility and user acceptance, offering explanations that users find both reasonable and informative. Objective evaluations further confirm its reliability, demonstrating high fidelity and faithfulness with competitive predictive performance.

Overall, the main contribution of CAM lies not in the concept extraction process itself, but in the mechanism that fuses textual semantics and quantitative reasoning into a unified, interpretable modeling framework. This enables CAM to deliver transparency without compromising accuracy, supporting trustworthy decision-making in high-stakes tabular data applications.

Future work

As a new attempt in the direction of combining knowledge mining and quantitative argumentation in interpretable research for tabular data, some aspects of this article are still preliminary. First, CAM only performs well in binary classification tasks, we will keep improving the modeling process to enhance its performance on multi-classification and regression tasks. Second, the explanation presented in this paper, based on quantitative argumentation, is interactive and has the advantage of being human-friendly due to its use of natural language. However, its generation method is mechanical. In future work, we will aim to integrate the framework of quantitative argumentation with generative artificial intelligence models to provide more flexible and human-like interactive explanations. Furthermore, our approach relies on textual descriptions of features for semantic knowledge mining. In real-world databases, these descriptions can be sparse, inconsistent, or even misleading, which may limit the quality of concept extraction and, consequently, model performance. Concurrently, to address the issue of model stability noted in our experiments, we plan to explore strategies for improving robustness while maintaining interpretability. For instance, we could incorporate techniques inspired by EBM, such as feature binning for continuous variables at the input layer. Lastly, extending the reasoning process to multi-relational argument types represents an open challenge and a promising direction for advancing the interpretability of tabular data models.

Data availability

This study utilizes four datasets: (1) the FICO dataset, publicly available at https://www.kaggle.com/c/home-credit-default-risk/data; (2) the MIMIC-III clinical database, accessible to qualified researchers upon completion of a data use agreement via https://physionet.org/content/mimiciii/1.4/; and (3–4) two proprietary datasets provided by Alibaba. The Alibaba data are confidential under corporate policies but available upon request through formal agreements. Contact Dr. Chi at haixiaochi@zju.edu.cn for access.

Received: 19 May 2025; Accepted: 25 November 2025
Published online: 08 January 2026

References

1. Minh, D., Wang, H. X., Li, Y. F. & Nguyen, T. N. Explainable artificial intelligence: A comprehensive review. *Artif. Intell. Rev.* **55**, 1–66 (2022).
2. Wan, M., Zha, D., Liu, N. & Zou, N. In-processing modeling techniques for machine learning fairness: A survey. *ACM Trans. Knowl. Discov. Data (TKDD)* **17**, 1–27 (2023).
3. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777 (2017).
4. Ribeiro, M. T., Singh, S. & Guestrin, C. " why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and data Mining*, 1135–1144 (2016).

5. Slack, D. Hilgard, S. Jia, E. Singh, S. & Lakkaraju, H. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI Ethics and Society*, 180–186 (2020).
6. Caruana, R. et al. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730 (2015).
7. Chang, C.-H. Caruana, R. & Goldenberg, A. NODE-GAM: Neural generalized additive model for interpretable deep learning. In *International Conference on Learning Representations* (2022).
8. Radenovic, F., Dubey, A. & Mahajan, D. Neural basis models for interpretability. *Adv. Neural Inf. Process. Syst.* **35**, 8414–8426 (2022).
9. Ibrahim, S., Afriat, G., Behdin, K. & Mazumder, R. GRAND-SLAMIN: Interpretable additive modeling with structural constraints. *Adv. Neural Inf. Process. Syst.* **36**, 61158–61186 (2024).
10. Čyras, K., Rago, A., Albini, E., Baroni, P. & Toni, F. Argumentative XAI: A survey. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence (IJCAI)* (2021).
11. Vassiliades, A., Bassiliades, N. & Patkos, T. Argumentation and explainable artificial intelligence: A survey. *Knowl. Eng. Rev.* **36**, e5 (2021).
12. Kampik, T., Čyras, K. & Alarcón, J. R. Change in quantitative bipolar argumentation: Sufficient, necessary and counterfactual explanations. *Int. J. Approx. Reason.* **164**, 109066 (2024).
13. Amgoud, L., Muller, P. & Trenquier, H. Leveraging argumentation for generating robust sample-based explanations. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence (IJCAI)*, 3104–3111 (2013).
14. Chi, H. & Liao, B. A quantitative argumentation-based automated explainable decision system for fake news detection on social media. *Knowl. Based Syst.* **242**, 108378 (2022).
15. Chi, H., Lu, Y., Liao, B., Xu, L. & Liu, Y. An optimized quantitative argumentation debate model for fraud detection in e-commerce transactions. *IEEE Intell. Syst.* **36**, 52–63 (2021).
16. Cocarascu, O., Rago, A. & Toni, F. Extracting dialogical explanations for review aggregations with argumentative dialogical agents. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 1261–1269 (Association for Computing Machinery, 2019).
17. Rago, A. & Toni, F. Quantitative argumentation debates with votes for opinion polling. In *International Conference on Principles and Practice of Multi-Agent Systems*, 369–385 (Springer, 2017).
18. Chen, C. et al. An interpretable model with globally consistent explanations for credit risk. arXiv preprint. arXiv:1811.12615 (2018).
19. Cayrol, C. & Lagasquie-Schiex, M.-C. On the acceptability of arguments in bipolar argumentation frameworks. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, 378–389 (Springer, 2005).
20. Potyka, N. Interpreting neural networks as quantitative argumentation frameworks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 6463–6470 (2021).
21. Rago, A., Toni, F., Aurisicchio, M. & Baroni, P. Discontinuity-free decision support with quantitative argumentation debates. In *Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning* (2016).
22. Eysenck, M. W. *Fundamentals of Cognition* 2nd edn. (Taylor and Francis, New York, 2012).
23. Gupta, M. & Agrawal, P. Compression of deep learning models for text: A survey. *ACM Trans. Knowl. Discov. Data (TKDD)* **16**, 1–55 (2022).
24. Murtagh, F. & Contreras, P. Algorithms for hierarchical clustering: An overview. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2**, 86–97 (2012).
25. Ghorbani, A., Wexler, J., Zou, J. & Kim, B. Towards automatic concept-based explanations. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 9277–9286 (2019).
26. Lee, S. et al. Self-explaining deep models with logic rule reasoning. In *36th Annual Conference on Neural Information Processing Systems* (2022).
27. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 (2016).
28. Kim, B. et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning*, 2668–2677 (PMLR, 2018).
29. Fang, Z., Kuang, K., Lin, Y., Wu, F. & Yao, Y.-F. Concept-based explanation for fine-grained images and its application in infectious keratitis classification. In *Proceedings of the 28th ACM International Conference on Multimedia*, 700–708 (2020).
30. Chen, C. et al. This looks like that: Deep learning for interpretable image recognition. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 8930–8941 (2019).
31. Marconato, E., Passerini, A. & Teso, S. Glancenets: Interpretable leak-proof concept-based models. *Adv. Neural Inf. Process. Syst.* **35**, 21212–21227 (2022).
32. Sheth, I. & Kahou, S. E. Auxiliary losses for learning generalizable concept-based models. *Adv. Neural Inf. Process. Syst.* **36**, 26966–26990 (2024).
33. Zarlenga, M. E. et al. Learning to receive help: Intervention-aware concept embedding models. *Adv. Neural Inf. Process. Syst.* **36**, 37849–37875 (2024).
34. Kim, E., Jung, D., Park, S., Kim, S. & Yoon, S. Probabilistic concept bottleneck models. In *International Conference on Machine Learning*, 16521–16540 (PMLR, 2023).
35. Poeta, E., Ciravegna, G., Pastor, E., Cerquitelli, T. & Baralis, E. Concept-based explainable artificial intelligence: A survey. arXiv preprint arXiv:2312.12936 (2023).
36. Tan, Z., Chen, T., Zhang, Z. & Liu, H. Sparsity-guided holistic explanation for LLMs with interpretable inference-time intervention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 21619–21627 (2024).
37. Barua, A., Widmer, C. & Hitzler, P. Concept induction using LLMs: A user experiment for assessment. In *International Conference on Neural-Symbolic Learning and Reasoning*, 132–148 (Springer, 2024).
38. Li, N. et al. CR-LLM: A dataset and optimization for concept reasoning of large language models. In *Findings of the Association for Computational Linguistics (ACL 2024)*, 13737–13747 (2024).
39. Adadi, A. & Berrada, M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* **6**, 52138–52160 (2018).
40. Chang, C.-H., Tan, S., Lengerich, B., Goldenberg, A. & Caruana, R. How interpretable and trustworthy are GAMs? In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 95–105 (2021).
41. Chen, J., Liao, K., Wan, Y., Chen, D. Z. & Wu, J. Danets: Deep abstract networks for tabular data classification and regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 3930–3938 (2022).
42. Panagiotou, E., Heurich, M., Landgraf, T. & Ntoutsi, E. TABCF: Counterfactual explanations for tabular data using a transformer-based VAE. In *Proceedings of the 5th ACM International Conference on AI in Finance*, 274–282 (2024).
43. Labaien, J. S., Uriguen, E. Z. & Garcia, X. D. C. Real-time, model-agnostic and user-driven counterfactual explanations using autoencoders. *Appl. Sci.* **13**, 2912 (2023).
44. Brughmans, D., Melis, L. & Martens, D. Disagreement amongst counterfactual explanations: How transparency can be misleading. *Top* **32**, 429–462 (2024).
45. Guidotti, R. Counterfactual explanations and how to find them: Literature review and benchmarking. *Data Min. Knowl. Discov.* **38**, 1–55 (2022).

46. Gajcin, J. & Dusparic, I. Redefining counterfactual explanations for reinforcement learning: Overview, challenges and opportunities. *ACM Comput. Surv.* **56**, 1–33 (2024).
47. Sendi, N., Abchiche-Mimouni, N. & Zehraoui, F. A new transparent ensemble method based on deep learning. *Procedia Comput. Sci.* **159**, 271–280 (2019).
48. Albini, E., Lertvittayakumjorn, P., Rago, A. & Toni, F. Deep argumentative explanations. arXiv preprint arXiv:2012.05766 (2020).
49. Vilone, G. & Longo, L. A global model-agnostic XAI method for the automatic formation of an abstract argumentation framework and its objective evaluation. In *Proceedings of the 1st International Workshop on Argumentation for eXplainable AI (ArgXAI, co-located with COMMA '22)*, 2119 (CEUR-WS, 2022).
50. Mollas, I., Bassiliades, N. & Tsoumakas, G. Truthful meta-explanations for local interpretability of machine learning models. *Appl. Intell.* (2023).
51. Prentzas, N., Pattichis, C. & Kakas, A. Explainable machine learning via argumentation. In *Explainable Artificial Intelligence xAI 2023* Vol. 1903 (ed. Longo, L.) (Springer, Cham, 2023).
52. Lakkaraju, H., Kamar, E., Caruana, R. & Leskovec, J. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 131–138 (2019).

Author contributions

Haixiao Chi and Beishui Liao conceived and designed the research study. Haixiao Chi drafted the manuscript and conducted the experiments in collaboration with Daiwei Wang. Haixiao Chi performed the data analysis and interpreted the results. Gaojie Cui and Feng Mao provided expert guidance on risk control and contributed to the experimental design. All authors reviewed and approved the final manuscript.

Funding

The research reported in this paper was partially supported by the Natural Science Foundation of Xiamen, China (No. 3502Z202474030), Fujian Provincial Natural Science Foundation of China (No. 2025J08322), National Natural Science Foundation of China (No. 62576309), and Alibaba Research Intern Program.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-30540-1.

Correspondence and requests for materials should be addressed to H.C. or B.L.