PeerJ Computer Science

CIAE: a consistency- and informativeness-aware explanation framework for improving type 2 diabetes prediction and mechanistic insights

Ning Wang[1,2], Wenchao Gu[1,2], Shuoming Liu[3,4], Minghui Wu[1,2], Chun Luo[5], Zhilei Chai[1,2] and Feng Zhang[3,4,6,7,8]

[1] School of Artificial Intelligence and Computer Science, Jiangnan College, Wu Xi, China
[2] Engineering Research Center of Intelligent Technology for Healthcare, Ministry of Education, Jiangnan College, Wu Xi, China
[3] School of Environment and Civil Engineering, Jiangnan University, Jiangnan College, Wu Xi, China
[4] Institute of Future Food Technology, JITRI, Jiangnan College, Yi Xing, China
[5] Shanghai Honsunbio Technology Co., Ltd, Shang Hai, China
[6] Department of Nutrition, Affliated Hospital of Jiangnan University, Jiangnan College, WuXi, China
[7] Wuxi School of Medicine, Jiangnan University, Jiangnan College, Wu Xi, China
[8] School of Biotechnology, Jiangnan College, Wu Xi, China

ABSTRACT

The gut microbiome plays a significant role in the development of type 2 diabetes (T2D), yet its underlying mechanisms remain unclear. Explainable artificial intelligence offers a promising approach to elucidating machine learning models and enhancing our understanding of these mechanisms. However, identifying the most effective explainer for specific problems remains a significant challenge. To address this issue, we propose the consistency- and informativeness-aware explanation framework (CIAE), which integrates multiple explainers to improve the consistency and informativeness of explanations. Using the top 100 Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthology (KO) features identified by CIAE, models achieved robust results on individual datasets (area under the curve (AUC): 0.870 and 0.808) and demonstrated strong cross-dataset generalization (AUC: 0.822 and 0.778). Notably, K00688 and K25026 were selected as key KOs in both datasets, and they are involved in the regulation of the starch and sucrose metabolism pathway and blood glucose homeostasis, which are closely related to the occurrence of diabetes.

Subjects Bioinformatics, Algorithms and Analysis of Algorithms, Artificial Intelligence, Data Mining and Machine Learning, Neural Networks

Keywords Gut microbiome, Type 2 diabetes, Explainable AI, KEGG orthology, Tabular deep learning

Submitted 17 March 2025
Accepted 27 August 2025
Published 27 October 2025

Corresponding author
Feng Zhang,
fengzhang@jiangnan.edu.cn

Academic editor
Davide Chicco

Additional Information and Declarations can be found on page 21

DOI **10.7717/peerj-cs.3223**

INTRODUCTION

Background

According to the International Diabetes Federation (IDF), the global prevalence of diabetes among adults have exceeded one-tenth in 2021 (*Sun et al., 2022*). Type 2 diabetes (T2D) accounts for the majority of diabetes cases, with over 90% (*Zheng, Ley & Hu, 2018*). Recent efforts have been converging towards discovering the association between the gut

microbiome and T2D (*Wu et al., 2020*; *Fan & Pedersen, 2021*; *Zhao et al., 2018*). Among them, the predominant emphasis has been on analyzing the associations between bacterial *taxa* and T2D. Some prior research have identified impact of bacterial *genera* on T2D through statistical or machine learning (ML) approaches using the data obtained from 16S rRNA sequencing (*Candela et al., 2016*; *Sedighi et al., 2017*; *Gou et al., 2021*). The 16S rRNA sequencing normally provides resolution at the genus level, however, highly diverse species could belong to the same genus (*Gurung et al., 2020*), leading to inconsistency in genus-level analysis across various studies regarding T2D. In contrast, shotgun metagenomic sequencing can unveil a broader diversity of species and the functional information (*Hugenholtz & Tyson, 2008*). In Metagenomic prediction Analysis based on Machine Learning (MteaML) (*Pasolli et al., 2016*), two types of features was extracted: species-level relative abundances and presence of strain-specific markers. Subsequently, a random forest (RF) model was used to identify the most important species for various disease prediction. More recently, deep learning (DL)-based approaches, including convolutional neural networks and autoencoders (*Li et al., 2021*; *Pinaya et al., 2020*), have emerged to predict the relationships between species and diseases (*Chen et al., 2022*; *Shen et al., 2022*). Traditional gene sequencing techniques are only able to identify specific species associated with disease. However, species are multi-dimensional, and their functions in different diseases are not consistent. Gut microbes usually participate in the body's metabolic activities through metabolites, and the production of metabolites is closely related to the function of enzymes carried by microorganisms. Therefore, the function performed by the gut microbiome may be related to the joint action of multiple species. Thus, it is one-sided to study the relationship between gut microbes and diseases only through a single species.

To address this issue, other studies attempt to discover the association between biological functions of gut bacteria with T2D by directly extracting functional biomarkers. The Non-supervised Orthologous Groups (eggNOG) (*Jensen et al., 2007*), the Kyoto Encyclopedia of Genes and Genomes (KEGG) (*Kanehisa & Goto, 2000*) and the Gene Ontology (GO) (*Ashburner et al., 2000*) are widely served as extensive knowledge bases of biological function. Several studies extracted function-related features, such as KEGG Orthology (KO) relative abundances through sequence alignment with the KEGG dataset. Further, conventional statistical approaches were applied to identify most relevant KOs for predicting T2D (*Qin et al., 2012*; *Karlsson et al., 2013*; *Dash et al., 2023*). Likewise, in the study (*Zhang et al., 2021*), the eggNOG and KEGG databases were used for gene annotation of microbiota. Incremental feature selection method were applied, followed by support vector machine (SVM) (*Hearst et al., 1998*) as the classification algorithm to predict T2D disease and identify latent biomarkers associated with T2D. In contrast, the *species* relative abundance and GO hierarchical structure information were utilized in *Liu, Zhang & Imoto (2021)* to predict T2D and liver cirrhosis diseases. Incorporating GO hierarchical structure information, the work extracted two types of GO features and the effectiveness both features were evaluated through extensive experiments. However, these prior studies either overlooked complex correlation and interaction among functional features or ignored their relative abundance differences between T2D and healthy control groups, potential leading to incomplete analysis in T2D.

Functional features, due to the intricate nature of microbial functions, are often represented as high-dimensional tabular data. Most of recent works on high-dimensional tabular data use either gradient-boosted decision trees (GBDT) based ML approaches, including eXtreme Gradient Boosting (XGBoost) (*Chen & Guestrin, 2016*), Light Gradient Boosting Machine (LightGBM) (*Ke et al., 2017*), and Categorical Boosting (CatBoost) (*Dorogush, Ershov & Gulin, 2018*), or deep neural networks, including Feature Tokenizer + Transformer (FT-Transformer) (*Gorishniy et al., 2021*), Attentive Interpretable Tabular Learning neural network (TabNet) (*Arik & Pfister, 2021*) and Weight Predictor Network with Feature Selection (WPFS) (*Margeloiu et al., 2023*). While GBDT approaches often outperform DL models in many scenarios, some state-of-the-art (SOTA) DL models have been reported to perform better than GBDT in various cases (*Gorishniy et al., 2021*; *Margeloiu et al., 2023*). Although MLs and DLs have achieved promising performance in disease prediction tasks, the lack of transparency and explainability of models' decisions poses a significant challenge. Currently, decision tree built-in feature importance, Local Interpretable Model-agnostic Explanations (LIME) (*Ribeiro, Singh & Guestrin, 2016*) and, SHapley Additive exPlanations (SHAP) (*Lundberg & Lee, 2017*) are widely used to identify important features in ML models (*Hancock & Khoshgoftaar, 2020*; *Li, 2022*; *Alabi et al., 2023*). In contrast, the *post-hoc* explanation methods, including Saliency (*Simonyan, Vedaldi & Zisserman, 2013*) SmoothGrad (*Smilkov et al., 2017*), and Integrated Gradients (*Sundararajan, Taly & Yan, 2017*) are employed to elucidate the decision-making process of DL models (*Dwivedi et al., 2024*; *Liang et al., 2023*). In reality, explanations provided by different explainers may vary, and it is challenging to identify a single explainer that performs best universally (*Agarwal et al., 2022*; *Roy et al., 2022*; *Krishna et al., 2022*). Different *post-hoc* explanation methods vary in their objectives, approximation domains and the loss functions they employ (*Han, Srinivas & Lakkaraju, 2022*; *Agarwal et al., 2022*). The suitability of *post-hoc* explanation methods varies with the type of input data across different tasks (*Han, Srinivas & Lakkaraju, 2022*). For example, continuous features are typically handled using additive continuous noise methods, such as SmoothGrad; binary features with binary noise methods, such as LIME; while most local function approximation-based methods are currently not applicable to discrete features. These fundamental differences collectively contribute to variations in the explanations generated by different explainers, with the quality of their results exhibiting clear distinctions across different tasks (*Roy et al., 2022*; *Krishna et al., 2022*). Therefore, selecting appropriate explainers or even designing specific explanation frameworks for specific tasks is particularly important.

Contribution

In this study, we propose consistency-and informativeness-aware explanation framework (CIAE), a novel explanation framework that effectively ensembles multiple explainers by considering both the consistency of explanations and the informativeness of the identified important features, to provide a reliable explanation on the most associated functional features in T2D prediction. Unlike traditional explainers that provide a single explanation result directly, CIAE starts from the perspective of the specific task and evaluates the

outputs of multiple explainers in terms of consistency and informativeness. By doing so, it integrates various explainers to derive more effective and reliable explanations. Our extensive experiments show that functional features directly extracted from metagenomic sequences are informative for T2D prediction. The top 100 KOs identified by CIAE show promising performance in T2D prediction tasks, both within a single T2D dataset and across multiple T2D datasets. Finally, we investigate how certain KO functional features play significant roles in T2D and whether they collectively participate in a pathway to influence the T2D.

Structure of the article

First, this article begins with an abstract that summarizes the overall content, followed by an introduction that provides background information and outlines the main contributions. Then, the methodology section includes data preprocessing and feature extraction, feature selection, the choice of a classifier for T2D, computing infrastructure, and assessment metrics. Next, this article presents the performance evaluation of the KO features, the aggregation of the results from interpretable models, the evaluation of the proposed interpretation optimization framework, and the evaluation of the significant KOs identified by the CIAE in the T2D dataset. Furthermore, this article also discusses the potential implications of the KOs identified through CIAE in the development of T2D. Finally, the article concludes with a discussion and summary of the findings.

MATERIALS AND METHODS

Data pre-processing and feature extraction

Our experiment analyzed two publicly available human gut microbiome datasets on T2D, including European female type 2 diabetes (EW-T2D) (*Karlsson et al., 2013*), and Chinese type 2 diabetes (C-T2D) (*Qin et al., 2012*). The two T2D datasets are shown in Table 1. C-T2D dataset contains total of 344 samples with 170 T2D samples and 174 control samples, while EW-T2D dataset contains total of 96 samples with 53 T2D samples and 43 control samples.

Our feature extraction pipeline is shown in Fig. 1. The initial inputs are raw metagenome sequencing reads obtained from two previous T2D cohorts. Subsequently, we apply a quality control step using fastp (*Chen et al., 2018*) to filter out low-quality reads and trim adapters. Additionally, human genomes are removed using Bowtie2 (*Langmead & Salzberg, 2012*), with the reference genome specified as Genome Reference Consortium Human Build 38 (GRCh38). The obtained clean reads are used to extract functional features using Diting which serves as a pipeline for generating KO relative abundances from the KEGG database. Within Diting (*Xue et al., 2021*), MEGAHIT (*Li et al., 2016*) is used to assembles the metagenomic reads and generate contigs, while Prodigal (*Hyatt et al., 2010*) is used to predict genes from these assembled contigs. Following that, Diting maps the input metagenomic reads to the predicted genes to obtain gene relative abundances, and generates KO relative abundances by aligning predicted proteins to the

Table 1 Human metagenomic datasets for T2D prediction.

Dataset	Total samples	Control samples	Case samples	KO features	Species features
EW-T2D	96	43	53	7,155	573
C-T2D	344	174	170	7,434	764

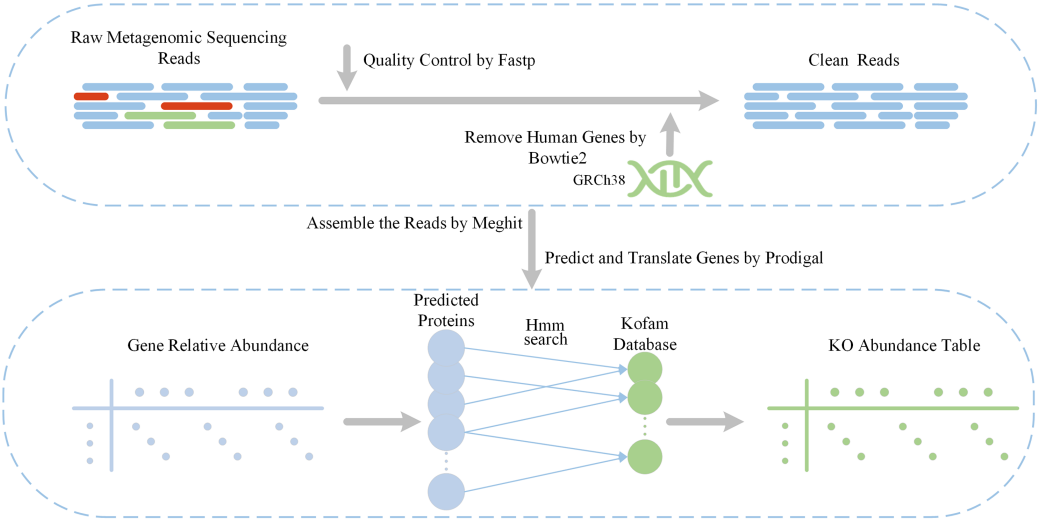

Figure 1 Data pre-processing pipeline and feature extraction.
Full-size DOI: 10.7717/peerj-cs.3223/fig-1

KOfam (*Aramaki et al., 2020*) database using hmmsearch (*Finn, Clements & Eddy, 2011*). Due to variations in sequence data across individual samples, the KO functional categories matched also differ, with potentially significant distributional differences between the experimental and control groups. For functional features that are not matched in certain samples, a value of 0 can be assigned, which does not affect the end-to-end predictive performance of the model. To address outliers such as excessively high feature abundances, all data must be normalized before being fed into the model to ensure stable and effective training.

Feature selection

Due to the high dimensionality and sparsity of the KO features, we apply feature selection based on the relative abundance and Pearson correlation coefficient.

Firstly, we filter out the KO features with relative abundances below 0.05% across all samples (*Boscaro et al., 2022*; *Sereti et al., 2023*), as calculated below:

$$\mu_j = f\left(\left[\frac{x_{1j}}{\sum_{i=1}^{d} x_{1i}}, \frac{x_{2j}}{\sum_{i=1}^{d} x_{2i}}, \ldots, \frac{x_{nj}}{\sum_{i=1}^{d} x_{ni}}\right]\right) \tag{1}$$

where $f(\vec{x})$ represents the maximum element in the vector, while u_j represents the maximum relative abundance value in the j-th KO feature. As mentioned in *Zhou, Wang & Zhu (2022)*, the higher the correlation between features (closer to 1), the greater the

redundancy among the features. Removing highly correlated features aims to eliminate redundant information.

Secondly, we calculated the correlation between features using the Pearson correlation coefficient and filtered out features with a correlation coefficient greater than 0.95. the correlation coefficient can be formulated as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2}$$

where x and y represent the abundance features in the samples after z-score normalization. The final KO features are selected if $\mu_j > 0.0005$ and $r < 0.95$. Through the aforementioned feature selection methods, we are able to eliminate irrelevant and redundant features in the samples. The number of KO features decreased from 7,155 to 1,188 in EW-T2D and from 7,434 to 1,813 in C-T2D, respectively.

T2D classification

As discussed previously, CatBoost and FT-Transformer are two primary artificial intelligence (AI) models in classification tasks using tabular data (*Borisov et al., 2022*; *Gorishniy et al., 2021*). In all experiments in this work, we apply these two classifiers.

CatBoost incorporates an innovative algorithm that automatically transforms categorical features into numerical features. Additionally, it employs an ordered boosting approach, which prevent bias in gradient estimation, resolve prediction shift issues, and minimize overfitting occurrences. The CatBoostClassifier is trained by minimizing the expected loss, which can be represented by the formula:

$$h^t = \arg\min_{h \in H} \mathbb{E}\mathscr{L}\left(y, F^{t-1}(\mathbf{x}) + h(\mathbf{x})\right) \tag{3}$$

where y represents the true output, $F^{t-1}(\mathbf{x})$ represents the prediction value of the first $t-1$ trees, and $h(\mathbf{x})$ represents the prediction of the current candidate decision tree.

FT-Transformer (*Gorishniy et al., 2021*) is a model designed for the tabular data, using the backbone of Transformer (*Vaswani et al., 2017*) architecture. It tokenizes features (categorical and numerical) into embeddings through random initialization and then feeds these embeddings into a Transformer architecture. The Binary cross entropy with Logits was utilized as the loss function:

$$\mathscr{L} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \cdot \log \sigma(x_i) + (1 - y_i) \cdot \log(1 - \sigma(x_i))\right] \tag{4}$$

where x_i represents the predicted value, y_i represents the true label, and σ represents the sigmoid activation function. The frameworks of the classifiers are shown in Fig. 2.

Explanation framework

Feature Importance in CatBoost's ensemble of decision trees is determined by assessing how frequently a feature is used for splitting and the consequent improvement in model performance resulting from these splits. *LIME* obtains local level explanation by training

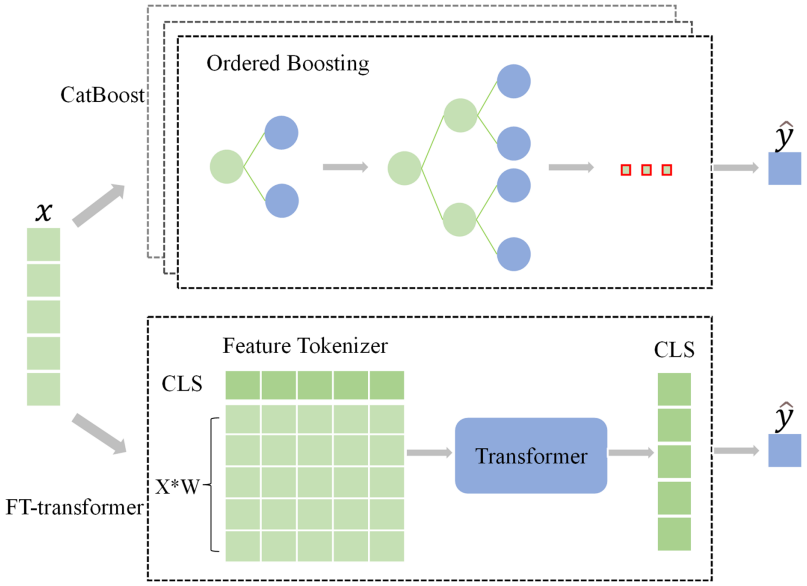

Figure 2 **The architecture of the classifiers.** Full-size DOI: 10.7717/peerj-cs.3223/fig-2

an interpretable model, such as a linear model, to gain the understanding of the behavior of a black-box model within a local neighborhood (*Ribeiro, Singh & Guestrin, 2016*). *TreeSHAP* is designed for interpreting and analyzing tree-based models (*Lundberg et al., 2020*). It captures the interaction effects among features, considering how features jointly influence predictions rather than affecting them in isolation. Compared to traditional Shapley values, which consider all possible subsets of features, TreeShap optimizes this by leveraging the tree structure (*Lundberg et al., 2020*). *Integrated Gradients* is guided by axioms and designed to attribute importance to features without requiring modifications to the original network, offering straightforward implementation. *Saliency* is previously used to interpret deep convolutional neural networks by highlighting the parts of an image that influence the class score. For tabular deep learning, it can be used to emphasize important features within tabular data. *DeepShap* is specifically tailored for interpreting deep learning models. It combines DeepLift (*Shrikumar, Greenside & Kundaje, 2017*), a method that backpropagates the contribution of all neurons in the network to each input feature by comparing each neuron's activation to a 'reference activation', and Shapley values. These explainers fall into different categories within the same local function approximation framework (*Han, Srinivas & Lakkaraju, 2022*). For example, some explainers combine continuous perturbations with gradient matching loss, while others use discrete perturbations with squared error loss. Each explainer is based on specific modeling assumptions and approximation strategies, which may introduce biases under certain data or task conditions. By integrating multiple explainers, their strengths can complement each other, reducing explanatory errors caused by the limitations of any single explainer.

In this study, we identify the features that have a significant impact on the model's overall prediction results by referring to the global explanation strategy of local function

approximation (*Han, Srinivas & Lakkaraju, 2022*), which involves taking the absolute values of multiple local explanation predictions and averaging the attribution scores. To mitigate the inconsistency in explanations provided by individual explainers, we apply an aggregation strategy to improve their reproducibility. Conversion of multiple trails' local explanations to aggregated global explanations can be expressed with the following formula:

$$I_i = \frac{1}{mn}\sum_{j=1}^{m}\sum_{k=1}^{n}\left|\left[w_{jk1}, w_{jk2}, \ldots, w_{jkd}\right]\right| \tag{5}$$

where $w_{jk} \in R^d$ represents a vector of attribution scores to each input feature for j-th trail's k-th sample and d represents the number of KO feature in T2D dataset. I_i represents the global explanation result of i-th explainer. To eliminate prediction biases across different explainers, it is important to standardize the global explanations provided by each explainer. This process standardizes the attribution scores within a consistent range, facilitating the integration of various explanations. The formula for the standardized global explanation of j-th feature from i-th explainer. f_{ij}, can be expressed as:

$$f_{ij} = \frac{I_{ij} - \bar{I}_i}{\sqrt{\frac{1}{d}\sum_{j=1}^{d}\left(I_{ij} - \bar{I}_i\right)^2}} \tag{6}$$

where I_{ij} represents the original attribution score of the j-th feature by the i-th explainer, and $\bar{I}_i$ represents the standard deviation of contribution scores of all features.

The consistency of explanations and informativeness of identified most important features are two essential dimensions for measuring the effectiveness of an explainer. Consequently, we propose the CIAE framework, an algorithm ensembles multiple model explainers by considering both of these factors. The formula is as follows:

$$g(x_j) = \frac{x_j}{\sum_{i=1}^{n} x_i}\left(y_{max} - y_{min}\right) + y_{min} \tag{7}$$

$$w = Softmax\left(\alpha\frac{1}{n}\sum_{i=1}^{n} g(R_i) + (1-\alpha)g(P)\right) \tag{8}$$

where R_i represents the performance of the explainers on the i-th reproducibility metric, and P represents the performance of the explainers on the area under the curve (AUC) metric. α represents the degree of importance placed on the reproducibility of explanations. Further, we normalize the evaluation metrics to the same range. $g(x_j) \in ((y_{min}, y_{max})$, and (y_{min}, y_{max}) defaulting to (0, 1) indicates the scaling range.

Subsequently, the final attribution score of the j-th feature is determined by the weighted sum of the attribution score assigned to the j-th feature by each explainer. The $score_j$ is formulated as:

$$score_j = \sum_{i=1}^{n} w_i f_{ij} \tag{9}$$

where f_{ij} represents the standardized attribution score of the i-th explainer for the j-th feature, and w_i represents the weight assigned to the i-th explainer based Eq. (8). Finally, the ranking of feature importance in *CIAE* can be expressed as:

$$CIAE = Rank(score, F) \tag{10}$$

where $score = [score_1, score_2, \ldots, score_d] \in R^d$ is a vector representing the attribution scores of all features, and $F = [F_1, F_2, \ldots, F_d] \in R^d$ is a vector representing the list of KO features. $Rank(x, y)$ represents sorting the *score* in descending order and mapping the indices to the F list to obtain the importance ranking of the KO features.

Computing infrastructure

All experiments were conducted on a system running Ubuntu 22.04. The deep learning models were implemented using PyTorch (version 1.12.1+cu116). The training and evaluation processes were performed on an NVIDIA A40 GPU to accelerate computation.

Assessment metrics

To comprehensively evaluate the performance of our classification models on the EW-T2D and C-T2D datasets, we employed five commonly used metrics: AUC, Accuracy (ACC), Recall, Precision, F1-score and Matthews Correlation Coefficient (MCC).

- AUC: AUC measures the area under the receiver operating characteristic (ROC) curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR). A higher AUC value indicates better discrimination between T2D and non-T2D samples, with 1.0 representing perfect classification and 0.5 indicating a random guess
- ACC: ACC is the proportion of correctly classified samples among all samples and is computed as:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

where *TP*, *TN*, *FP*, and *FN* represent the number of True Positives, True Negatives, False Positives, and False Negatives, respectively.

- Recall: recall measures the ability of the model to correctly identify positive samples (T2D cases) and is given by:

$$Recall = \frac{TP}{TP + FN}. \tag{12}$$

A higher recall indicates fewer missed positive cases.

- Precision: precision quantifies the proportion of correctly predicted positive cases among all predicted positives:

$$Precision = \frac{TP}{TP + FP}. \tag{13}$$

High precision implies that most of the samples classified as T2D are indeed positive cases, reducing false alarms.

- F1-score: F1-score is the harmonic mean of Precision and Recall, balancing both metrics in cases where one is disproportionately high or low:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{14}$$

A high F1-score indicates a well-performing model, especially in datasets with class imbalance.

- MCC: MCC is used to evaluate the performance of binary classification models, particularly suitable for imbalanced class scenarios. It ranges from −1 to 1, where −1 indicates completely incorrect prediction, 0 indicates prediction no better than random guessing, and 1 indicates perfect prediction.

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{15}$$

A higher MCC value indicates better discrimination between T2D and non-T2D samples.

RESULTS

Performance evaluation

In this study, we adapted Feature Tokenizer + Transformer (FT-Transformer) (*Gorishniy et al., 2021*) and Categorical Boosting (CatBoost) (*Dorogush, Ershov & Gulin, 2018*) as our classifiers. Both models are are respectively the SOTA models in DL and ML for tabular data. During the experiments, we randomly split the dataset into 80% training and 20% testing sets. Further, within the training dataset, we reserved 20% of the data as the validation set to prevent overfitting. To increase the reproducibility of the results, we conducted five trials using different random seeds, while maintaining the same hyperparameters. Subsequently, we presented the averaged outcomes, ensuring a robust and reliable evaluation. We evaluated performance using the following five metrics: AUC, ACC, Recall, Precision and F1-score. The evaluation results for EW-T2D and C-T2D dataset are shown in Table 2. On the EW-T2D dataset, CatBoost using KO features outperform other model in terms of all five metrics, achieving an AUC of 0.830. While on the C-T2D dataset, CatBoost using KO features receive the best AUC of 0.777, the best Acc, Recall, F1-score are achieved by FT-Transformer using KO features with values of 0.715, 0.735, and 0.717 respectively. The best precision is achieved by both CatBoost and FT-Transformer on KO features, both with a precision of 0.702. Overall, on both datasets, models using KO relative abundance features outperform models using species relative abundance features, which evidence that KO relative abundance features are essential informative features in T2D prediction.

Model aggregation

In the field of model explainability, there is a lack of a common foundational approach and consistent conceptual or practical understanding of explanations. It is necessary to choose appropriate explainers for specific models and scenarios (*Han, Srinivas & Lakkaraju, 2022*). CatBoost, a type of GBDT-based approach, commonly uses LIME, TreeSHAP

Table 2 Performance evaluations on EW-T2D and C-T2D datasets.

Dataset	Feature	Methods	AUC	ACC	Recall	Precision	F1
EW-T2D	Species	FT-Transformer	0.783	0.680	0.690	0.713	0.699
		CatBoost	0.806	0.730	0.745	0.754	0.742
	KO	FT-Transformer	0.810	0.750	0.818	0.756	0.782
		CatBoost	**0.830**	**0.770**	**0.836**	**0.772**	**0.799**
C-T2D	Species	FT-Transformer	0.738	0.663	0.676	0.670	0.666
		CatBoost	0.745	0.681	0.652	0.685	0.665
	KO	FT-Transformer	0.765	**0.715**	**0.735**	**0.702**	**0.717**
		CatBoost	**0.777**	0.707	0.705	**0.702**	0.703

Note:
The best performance is in **bold**.

(*Lundberg et al., 2020*) and its inherent feature importance to provide model explanation (*Hancock & Khoshgoftaar, 2020*; *Li, 2022*; *Alabi et al., 2023*). In contrast, DL-based models like FT-Transformer normally adapt Integrated Gradients, Saliency, and DeepSHAP (*Lundberg & Lee, 2017*) to provide reliable explanations (*Dwivedi et al., 2024*; *Liang et al., 2023*). We adhere to the convention in this work. Furthermore, we assessed the reproducibility of the behavior of an individual explainer by comparing the overlapping and Spearman's correlation of top 100 features identified by that explainer between pairs of trials. The average consistency results over all pairs of trials, represented by blue bars, are reported in Fig. 3. Subsequently, we introduced model aggregation, which was also used in *Zhao, Shao & Asmann (2022)*, to average the contribution scores of features obtained through an explainer across multiple trails. The overall consistencies were calculated by averaging the consistency scores between the aggregated results and the results from each trail. They are represented by green bars in Fig. 3. The results indicate that the model aggregation can effectively reduce explanation biases cross different trails, thus improve the reproducibility of model's explanation.

Subsequently, we examined the detailed consistency across different classifiers and their corresponding explainers. *Inter Classifiers:* the overall consistency of explainers on Catboost classifier outperforms that on FT-Transformer. This may be attributed to the fact that the FT-Transformer has a significantly more intricate architecture compared to Catboost, resulting in a more complex explanation. A slight variation in model parameters could lead to notable fluctuations in the explanation results. *Intra Classifier:* Integrated Gradients, Saliency, and DeepShap explainers show relatively similar consistency in terms of Spearman's correlation and overlapping score on C-T2D set. However, Saliency demonstrates better consistency in terms of Spearman's correlation and overlapping score on EW-T2D set. In terms of Spearman's correlation, TreeShap shows best consistency on EW-T2D set, while CatBoost exhibits highest consistency on C-T2D set. However, in terms of overlapping score, both Lime and TreeShap show higher consistency over CatBoost. The results indicate that there is no universal explainer that consistently performs best across different classifiers and consistency evaluation metrics. Consequently,

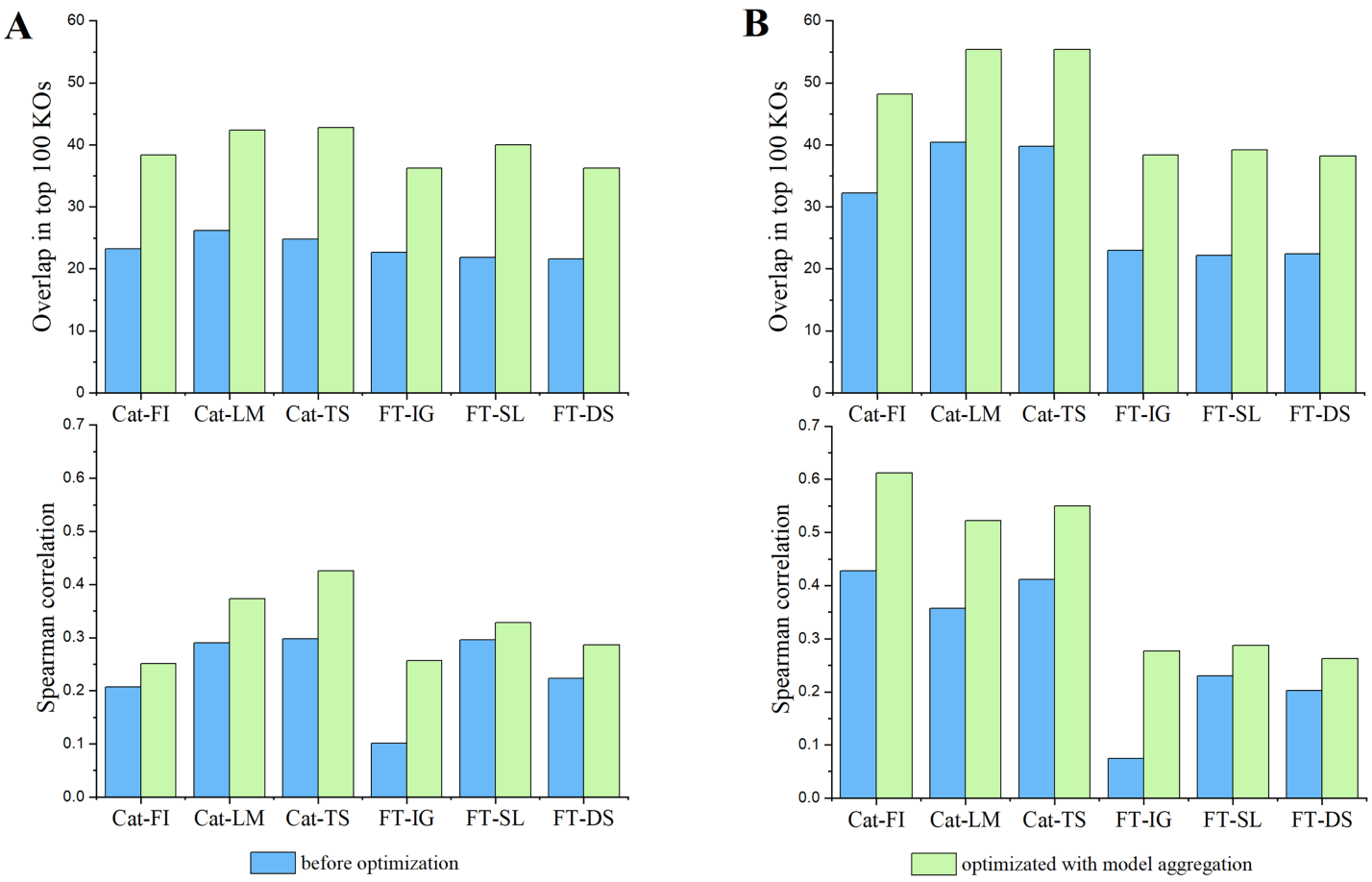

Figure 3 The impact of model aggregations on model reproducibility. Classifiers: **Cat**-* represents CatBoost; **FT**-* represents FT-Transformer. Explanation methods: FI represents Feature importance; LM represents LIME; TS represents TreeSHAP; IG represents Integrated Gradients; SL represents Saliency; DS represents DeepSHAP. (A) Overlaps among the top 100 significant KOs across various classifiers and explanation methods (upper panel) and spearman's correlation of KO contribution scores (lower panel) in EW-T2D dataset. (B) Same as (A) but on C-T2D dataset.
Full-size DOI: 10.7717/peerj-cs.3223/fig-3

for the remainder of the article, we use the aggregated explanation over five trials to represent the explanation reported by an explainer.

Optimized explanation framework

As analyzed in the previous section, despite significant successes in interpretable AI (IAI) and explainable AI (XAI), a common foundational and consistent approach is still lacking (*Agarwal et al., 2022*; *Han, Srinivas & Lakkaraju, 2022*). The main reason behind is that individual explainers struggle to capture all characteristics and underlying reasons of a model's decisions (*Roy et al., 2022*; *Krishna et al., 2022*). In this context, finding a reliable approach to effectively integrate various explainers into a unified framework becomes an important topic. To achieve that, we propose a consistency and informativeness aware explanation (CIAE) framework, which effectively ensembles multiple explanations to reduce the bias caused by individual explainers. In Fig. 4, we show the overlap of the top 100 significant KOs obtained from each explainer on two T2D dataset. The results indicate

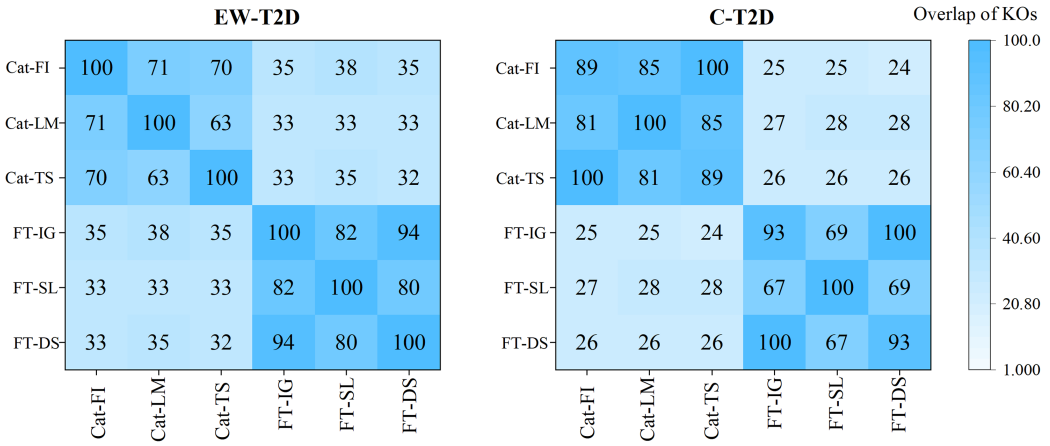

Figure 4 Overlaps among the top 100 significant KOs across various classifiers and explanation methods. Full-size DOI: 10.7717/peerj-cs.3223/fig-4

that the degree of overlap is significantly higher in the intra-classifer explanations compared with inter-classifer explanations. This disparity may be due to the different architectures of CatBoost and the FT-Transformer, which result in distinct predictive patterns and preferences (*Dargan et al., 2020*). Within a single classifier, there is still some variation across different datasets. However, the average overlap is above 70%, indicating that these explainers agree on the majority of top significant KOs while disagreeing on some of them. This result is more clearly displayed in the Venn diagram in Fig. 5. Afterward, we assessed the informativeness of KOs chosen by each explainer. Table 3 shows the evaluation results on EW-T2D and C-T2D datasets using top 100 KOs chosen by individual explainers and the CIAE framework. We started with comparing the effectiveness among individual explainers. On the EW-T2D dataset, using FT-Transformer as the classifier, the highest AUC of 0.858 is achieved with the top 100 KOs identified by Saliency. In contrast, using CatBoost as the classifier, the highest AUC of 0.868 is achieved with the top 100 KOs identified by TreeShap. On the C-T2D dataset, using FT-Transformer as the classifier, the highest AUC of 0.802 is achieved with the top 100 KOs identified by Integrated Gradients. In contrast, using CatBoost as the classifier, the highest AUC of 0.805 is achieved with the top 100 KOs identified by Feature Importance. On average, the AUC improved by 3% to 4% when using 100 KOs identified by the explainers. This might be due the existence of numerous irrelevant or redundant features within high-dimensional inputs. IAI and XAI-based explainers can effectively identify the most informative features and efficiently reduce data dimensionality, thus improving the classifier's performance. This finding is consistent with the results reported in the work (*Liu et al., 2022*). Consequently, we propose the CIAE framework, which takes into consideration both the consistency of explanations and the informativeness of the identified important features to optimize the overall explanations of classifier's behavior. The details of CIAE framework is described in Explanation framework. As shown in Table 3, our CIAE framework outperforms all other explainers across all T2D datasets and classifiers.

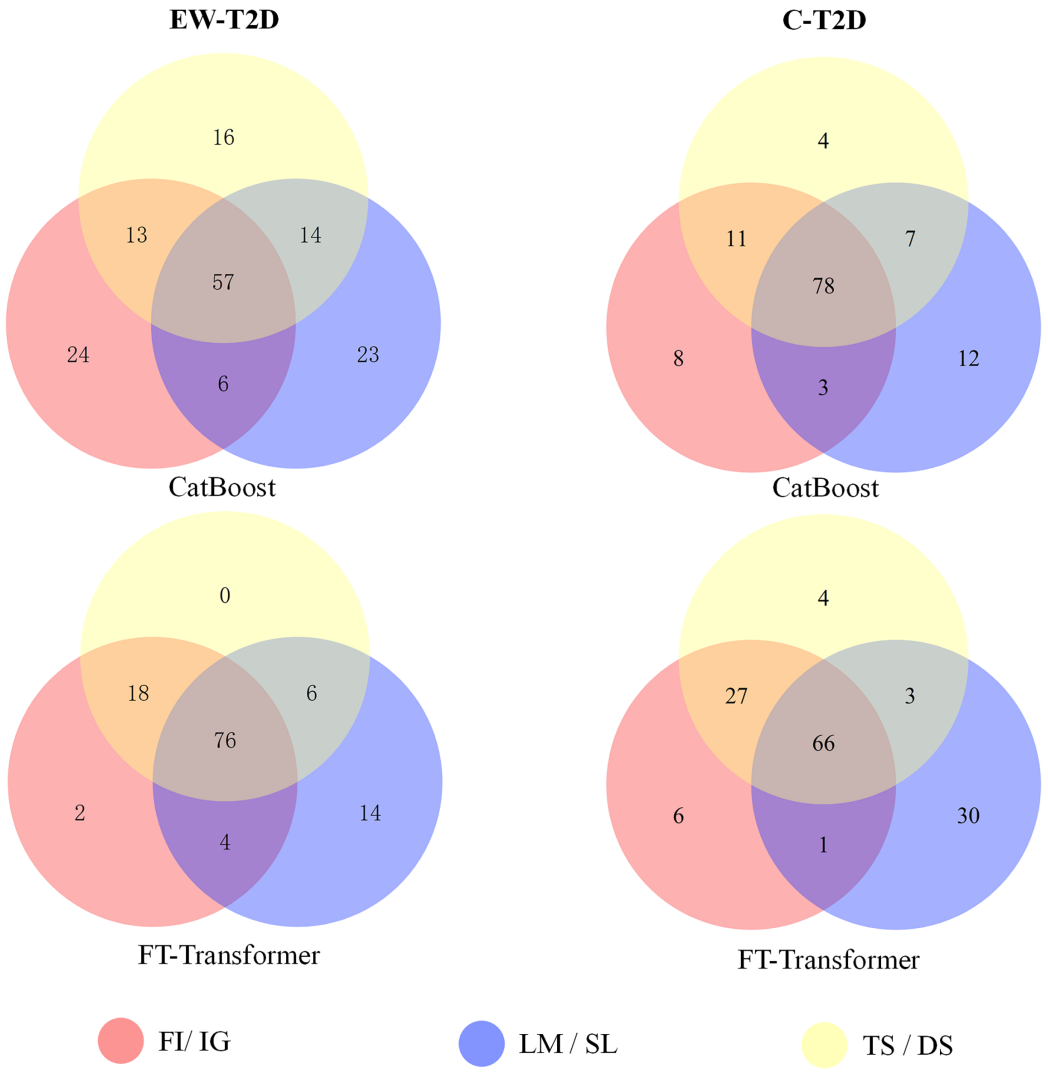

Figure 5 **The Venn diagram illustrates the overlaps of the top 100 significant KOs identified by various explainers using CatBoost and FT-Transformer as the classifiers.**

Full-size DOI: 10.7717/peerj-cs.3223/fig-5

Evaluation of significant KOs identified by CIAE across T2D datasets

In gut microbiome research, cross-dataset validation of important features is crucial for identifying disease-associated biomarkers (*Dai et al., 2022*). *Li et al. (2023)* conducted a cross-cohort validation study on 20 diseases using machine learning models and analyzed the consistency of disease biomarkers across different cohorts for the same diseases. *Gao et al. (2024)* developed xMarkerFinder, a four-stage computational framework for identifying and validating microbial biomarkers across cohorts. By integrating meta-analysis, machine learning, and biomarker interpretation methods, the framework efficiently identifies robust and highly specific biomarkers. In cross-dataset studies of hepatitis C, machine learning and pre-clustering techniques were used to improve detection accuracy, while XAI tools enhanced model interpretability, helping medical

Table 3 AUC scores for T2D prediction using the top 100 significant KOs identified by different explainers.

Dataset	Classifier	Explainer	AUC
EW-T2D	FT-Transformer	FT-IG	0.842
		FT-SL	0.858
		FT-DS	0.848
		FT-CIAE	**0.870**
	CatBoost	Cat-FI	0.854
		Cat-LM	0.864
		Cat-TS	0.868
		Cat-CIAE	**0.869**
C-T2D	FT-Transformer	FT-IG	0.802
		FT-SL	0.801
		FT-DS	0.797
		FT-CIAE	**0.807**
	CatBoost	Cat-FI	0.805
		Cat-LM	0.794
		Cat-TS	0.801
		Cat-CIAE	**0.808**

Note:
The best performance is in **bold** and the second best performance is underlined.

professionals better understand and trust the model's predictions (*Sharma, Khade & Satapathy, 2025*).

In this section, we evaluate the effectiveness and generalization of the top significant KOs identified by CIAE framework across two T2D datasets. Initially, we identified the top 100 significant KOs within each T2D dataset using individual explainers and our CIAE approaches. Subsequently, we tested the generalization of these top significant KOs across T2D datasets. More specifically, we trained CatBoost and FT-Transformer on the C-T2D dataset using the top 100 significant KOs identified by the explainers from the EW-T2D dataset, and vice versa. The results are shown in Table 4. On the C-T2D dataset, the best AUC score of 0.7789 is achieved with CatBoost as the classifier and Cat-CIAE as the explainer. This represents a performance improvement of approximately 0.2% compared to directly using CatBoost on the C-T2D dataset with all KOs. Additionally, when using FT-Transformer as the classifier, the best AUC score of 0.7508 is achieved with FT-CIAE as the explainer. However, the performance dropped by approximately 1.5% compared to directly FT-Transformer on the C-T2D dataset with all KOs. On EW-T2D dataset, the best AUC score of 0.8222 is achieved with FT-Transformer as the classifier and FT-CIAE as the explainer. This represents a performance improvement of approximately 1.2% compared to directly using FT-Transformer on the EW-T2D dataset with all KOs. Additionally, when using CatBoost as the classifier, the best AUC score of 0.7980 is achieved with Cat-CIAE as the explainer. However, the performance dropped by approximately 3% compared to directly CatBoost on the EW-T2D dataset with all KOs, the results are shown in Table 5. Meanwhile, an analysis of our CIAE approaches and other

Table 4 AUC scores for T2D prediction using the top 100 significant KOs identified by different explainers across T2D datasets.

Dataset		C-T2D		EW-T2D	
Classifier		**FT-Transformer**	**CatBoost**	**FT-Transformer**	**CatBoost**
Single explanation	Cat-FI	0.7063	0.7389	0.7858	0.7697
	Cat-lM	0.7226	0.7670	0.7374	0.7697
	Cat-TS	0.7005	0.7512	0.7656	0.7778
	FT-IG	0.7279	0.7319	0.8000	0.7899
	FT-SL	0.7279	0.7522	0.7515	0.6909
	FT-DS	0.7361	0.7474	0.7818	0.7818
CIAE explanation	Cat-CIAE	0.7418	**0.7789**	0.8121	**0.7980**
	FT-CIAE	**0.7508**	0.7559	**0.8222**	0.7858

Note:
The best performance is in **bold.**

Table 5 MCC scores for T2D prediction using the top 100 significant KOs identified by different explainers across T2D datasets.

Dataset		C-T2D		EW-T2D	
Classifier		**FT-Transformer**	**Catboost**	**FT-Transformer**	**CatBoost**
Single explanation	Cat-FI	0.3567	0.3753	0.2244	0.4817
	Cat-lM	0.3420	0.3852	0.3117	0.5065
	Cat-TS	0.2513	0.3501	0.2921	0.4923
	FT-IG	0.3573	0.3226	0.4897	0.3968
	FT-SL	0.3070	0.3720	0.2539	0.3440
	FT-DS	0.3281	0.3651	0.4100	0.4191
CIAE explanation	Cat-CIAE	0.3741	**0.4003**	0.3708	**0.5483**
	FT-CIAE	**0.3883**	0.3758	**0.5347**	0.4265

Note:
The best performance is in **bold**.

individual explainers stability concerning the receiver operating characteristic (ROC) curve is shown in Fig. 6. In addition, we further employed MCC as an additional metric. The results are shown in Table 5. On the C-T2D dataset, the best MCC score of 0.4003 was achieved using CatBoost as classifier and Cat-CIAE as the explainer, while using FT-Transformer with FT-CIAE yielded the best MCC score of 0.3883. On the EW-T2D dataset, the best MCC score of 0.5483 was achieved using CatBoost as classifier and Cat-CIAE as the explainer, while using FT-Transformer with FT-CIAE yielded the best MCC score of 0.5347.

We conclude our findings as follows:

- The top 100 significant KOs show a higher ability to generalize when the source and target classifiers are identical. As previously discussed, this phenomenon may be due to the different architectures of CatBoost and FT-Transformer, which lead to distinct predictive patterns and preferences (*Dargan et al., 2020*).

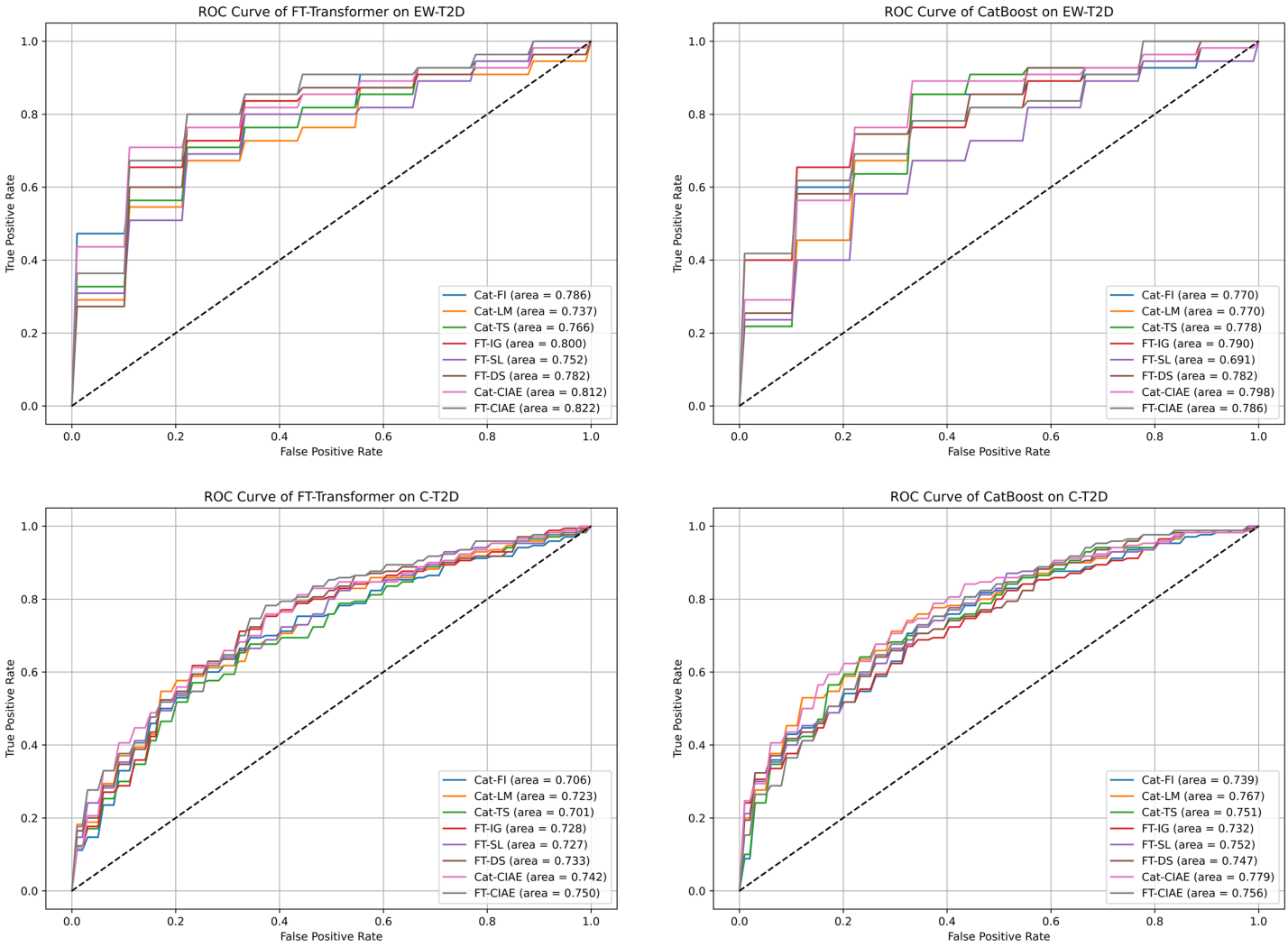

Figure 6 **ROC curves and AUCs for CIAE and baseline explainers we experimented with.** Full-size DOI: 10.7717/peerj-cs.3223/fig-6

- Compared to other explainers, Cat-CIAE and FT-CIAE consistently outperform them, regardless of whether the source and target classifiers are identical or not. This indicates that our proposed CIAE framework is more robust and generalizable compared to existing single -explainer approaches.
- Given the promising performance of our proposed CIAE framework in T2D prediction tasks, both within a single T2D dataset and across multiple T2D datasets, it has the potential to serve as a efficient and reliable feature selection tool to mitigate overfitting issues commonly exist in high-dimensional, low sample-size dataset.

Underlying significance of identified KOs in the development of T2D

In this section, we discuss and validate the most significant KOs identified by our proposed CIAE explainer in the context of T2D. Detailed information of those KOs is provided in Table 6.

Table 6 The KOs associated with T2D from KEGG.

KO	Function	Reference	The role in T2D	Dataset
K00688	Glycogen phosphorylase	*Alruhaimi et al. (2023)*	Controls blood sugar by hydrolysis of starch to produce glucose-1P.	EW-T2D C-T2D
K25026	Glucokinase	*Li & Zhu (2024)*	The commander in the glucose homeostasis regulatory system promptly regulates the secretion of the blood-glucose regulating hormones-insulin and glucagon.	EW-T2D C-T2D
K07029	Diacylglycerol kinase (ATP)	*Massart & Zierath (2019)*	Diacylglycerol accumulation is associated with the development of lipid-induced insulin resistance.	C-T2D
K01610	Phosphoenolpyruvate carboxykinase (ATP)	*Gonzalez-Rellan et al. (2023)*	Controls glucose metabolism.	C-T2D
K21836	D-lactate dehydrogenase (acceptor)	*Lam et al. (2005)*	Hypothalamic pyruvate metabolism regulates blood glucose, potentially improving glucose homeostasis in diabetic patients.	C-T2D
K01687	Dihydroxy-acid dehydratase	*Jang et al. (2016)*	Activation of branched-chain amino acid (BCAA) metabolism improves glucose and lipid homeostasis.	EW-T2D
K01658	Anthranilate synthase component II	*Ruiz-Canela et al. (2018)*	Aromatic amino acids (AAAs): phenylalanine, tyrosine, and tryptophan were positively associated with T2D risk	EW-T2D
K00950	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine diphosphokinase	*Lind et al. (2019)*	Folate supplementation may be beneficial for glucose homeostasis and reducing insulin resistance.	EW-T2D
K03465	Thymidylate synthase (FAD)	*Yamada, Mendoza & Koutmos (2023)*	5-Formyltetrahydrofolate promotes the conformational reshaping of the active site of methylenetetrahydrofolate reductase and inhibits its activity.	EW-T2D

The prominent functions obtained from EW-T2D

Among the top 10 significant KOs obtained in the EW-T2D dataset, K01687 and K01658 are involved in the biosynthesis of amino acids and are associated with the synthesis of valine, isoleucine and tryptophan. These three amino acids, or their intermediates are associated with the occurrence and development of diabetes (*Jang et al., 2016*; *Menni et al., 2013*; *Gao et al., 2023*). In addition, The involvement of K03465 and K00950 in the synthesis of folic acid is also noteworthy. Folic acid can improve insulin resistance in diabetic patients, and can play a role in alleviating diabetes (*Yang et al., 2023*).

The prominent functions obtained from C-T2D

Similarly, among the top 10 significant KOs obtained from C-T2D dataset, K01610 and K21836 are involved in the Pyruvate metabolism pathway and play important roles in energy metabolism. K07029 is involved in the Glycerolipid metabolism pathway, disorders of which can cause insulin resistance in the body, which in turn can lead to obesity and diabetes.

The prominent functions obtained from both EW-T2D and C-T2D

Among the top 10 of important KOs, K00688 and K25026, identified from both datasets, have been shown to have a strong association with T2D in previous studies. K00688 and K25026 are involved in the Starch and sucrose metabolism pathway, as shown in Fig. 7. The K00688 encodes the synthesized glycogen phosphorylase that catalyzes the formation of starch to D-Glucose 1-phosphate, which further generates D-Glucose. The K25026

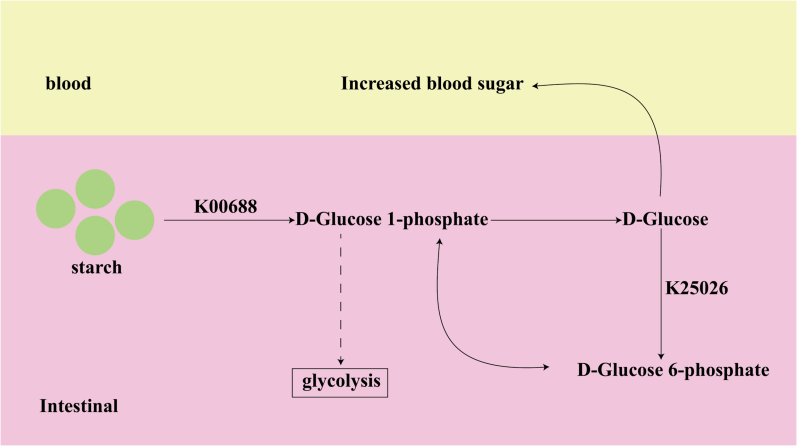

Figure 7 **The function of key biomarkers.** Full-size DOI: 10.7717/peerj-cs.3223/fig-7

encodes the synthesized glucokinase that catalyzes the formation of D-Glucose 6-phosphate, which can be interconverted with D-Glucose 1-phosphate, and they are involved in the glycolytic pathway. These two KOs not only affect the progress of subsequent glycolysis, but their intermediate metabolite D-Glucose will also directly cause an increase in blood glucose, so they are closely related to the occurrence and development of diabetes.

DISCUSSION

Numerous earlier studies have evaluated and delineated potential biomarkers concerning T2D using gut microbiota data. However, most of these studies primarily analyze at taxonomic levels and overlook the functional and systematic exploration (*Chen et al., 2022*; *Shen et al., 2022*). Through extensive experiments in this study, we observed that functional features, such as KO relative abundances, outperformed taxonomic features, such as species-level relative abundances, in T2D prediction, achieving the highest AUC scores of 0.830 and 0.777 for EW-T2D and C-T2D, respectively. In addition, compared to studying the association between species and the T2D, directly analyzing the function of gut microbiota provides valuable insights into the role of these microorganisms in the onset and progression of T2D. With the help of machine learning and XAI techniques, researchers can gain deeper insights into the pathogenesis of T2D (*Chari et al., 2023*; *Tuppad & Patil, 2024*; *Sharma, Saini & Hasija, 2024*). However, due to inconsistency in the explanation results, there may be deviations or even misinterpretations in understanding the mechanisms underlying T2D. Furthermore, to identify the most significant KO biomarkers in T2D prediction, we introduced a novel explanation framework, CIAE, which can effectively ensembles various explainers by considering both the consistency and informativeness provided by each explainer. CIAE demonstrates promising performance on both single T2D datasets and across multiple T2D datasets. Extensive experiments indicated that using the top 100 identified KOs for T2D prediction received AUC scores of 0.870 on the EW-T2D dataset and 0.808 on the C-T2D dataset. These KOs

continued to demonstrate competitiveness, achieving AUC scores of 0.822 and 0.789 in predicting T2D across two distinct T2D datasets.

Finally, in our biomarker discussion, we have predicted the KOs that are related to diabetes through artificial intelligence, which provides a new research direction for us to understand and treat diabetes. The results showed that K00688 and K25026 were selected as key KOs in both datasets, they were involved in the metabolism of starch and the regulation of blood glucose homeostasis in the body, and were closely related to the occurrence of diabetes. K01610, K21836, and K07029 were screened from the C-T2D dataset for the metabolism of carbohydrates and lipids. K01687, K01658, K03465 and K00950 were mainly related to amino acid metabolism in the EW-T2D dataset. This may be related to differences in eating habits, which suggests that the causes of diabetes in different regions are different, and the means to prevent diabetes should also be adapted to local conditions.

CONCLUSIONS

This study analyzed the relationship between functional features (KO features) of the gut microbiota and T2D to improve prediction and interpretation. Compared to traditional taxonomic features (species features), functional features demonstrated superior predictive ability for T2D, achieving the highest AUC scores of 0.830 and 0.777 on the EW-T2D and C-T2D datasets, respectively. These findings suggest that directly analyzing the functional characteristics of the gut microbiota provides more insight into their roles in the onset and progression of T2D. Furthermore, the proposed consistency- and informativeness-aware explanation framework (CIAE) effectively integrates multiple explainers, significantly improving the consistency and informativeness of model explanations. The CIAE framework demonstrated promising performance in both single-dataset and cross-dataset experiments. Using the top 100 significant KO features identified by CIAE, the models achieved competitive results on individual datasets (EW-T2D: AUC = 0.870, C-T2D: AUC = 0.808) and exhibited strong generalization capabilities in cross-dataset T2D prediction (EW-T2D: AUC = 0.822, C-T2D: AUC = 0.778). These results highlight the potential of the CIAE framework as an efficient and reliable feature selection tool for high-dimensional, low-sample-size datasets.

Through biological analysis of the most significant KO features, we identified K00688 and K25026 as key KOs in both T2D datasets, which are closely associated with starch metabolism and blood glucose homeostasis. Subsequently, K01610, K21836, and K07029, identified in the C-T2D dataset, were primarily involved in carbohydrate and lipid metabolism, while K01687, K01658, K03465, and K00950, identified in the EW-T2D dataset, were associated with amino acid metabolism. These differences may be related to regional eating habits, suggesting that the mechanisms underlying T2D can vary between populations, and prevention strategies should be tailored to local conditions. In conclusion, this study not only validated the importance of functional features in the prediction of T2D, but also introduced the CIAE framework as a novel approach to feature selection and model interpretation in high-dimensional data. The identified key KOs

provide valuable references for further research into the mechanisms of T2D and potential therapeutic targets.

Due to limited access to metagenomic samples of T2D, the conclusions drawn in this study might be insufficient and may contain biases. On the one hand, due to differences in sequencing methods, metagenomic data may exhibit systematic biases across datasets. Additionally, limited coverage of the KEGG database often leads to incomplete annotation of KO functions. These factors may collectively undermine the stability and generalizability of the model in cross-dataset validation. On the other hand, the high-dimensional nature of KO features poses a significant challenge for algorithms in making accurate predictions. In our future work, we will explore larger and more diverse T2D sample sets while integrating various data augmentation approaches to expand microbial datasets. Additionally, exploring dimension reduction approaches such as autoencoders and PCA could be a promising direction to improve prediction performance for T2D.

Moreover, the essential future direction explores deep learning models specifically designed for biomedical tabular data, which are not only suitable for high-dimensional, small-sample datasets, but also possess interpretability, thereby facilitating the end-to-end identification of key KOs related to T2D. This would enable the rapid and efficient discovery of important functional biomarkers across multiple T2D datasets.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Funding provided from grants of the National Key Research and Development Program of China (2022YFF1100601), the National Natural Science Foundation of China (82370809), and Fundamental Research Funds for the Central Universities (JUSRP123035). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Key Research and Development Program of China: 2022YFF1100601.
National Natural Science Foundation of China: 82370809.
Fundamental Research Funds for the Central Universities: JUSRP123035.

Competing Interests

Chun Luo is an employee of the Shanghai Honsunbio Technology Co., Ltd.

Author Contributions

- Ning Wang conceived and designed the experiments, performed the experiments, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

- WenChao Gu conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Shuoming Liu analyzed the data, prepared figures and/or tables, and approved the final draft.
- Minghui Wu performed the computation work, prepared figures and/or tables, and approved the final draft.
- Chun Luo analyzed the data, prepared figures and/or tables, and approved the final draft.
- Zhilei Chai conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Feng Zhang conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The code is available at GitHub and Zenodo:

- https://github.com/gwcde/CIAE.

- gwcde. (2025). gwcde/CIAE: main (main). Zenodo. https://doi.org/10.5281/zenodo.17064003.

The EW-T2D dataset is available at European Nucleotide Archive (ENA): ERP002469/PRJEB1786.

The C-T2D dataset is available at the National Center for Biotechnology Information (NCBI): SRA045646.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.3223#supplemental-information.

REFERENCES

Agarwal C, Krishna S, Saxena E, Pawelczyk M, Johnson N, Puri I, Zitnik M, Lakkaraju H. 2022. OpenXAI: towards a transparent evaluation of model explanations. *Advances in Neural Information Processing Systems* **35**:15784–15799.

Alabi RO, Elmusrati M, Leivo I, Almangush A, Mäkitie AA. 2023. Machine learning explainability in nasopharyngeal cancer survival using LIME and SHAP. *Scientific Reports* **13(1)**:8984 DOI 10.1038/s41598-023-35795-0.

Alruhaimi RS, Mostafa-Hedeab G, Abduh MS, Bin-Ammar A, Hassanein EH, Kamel EM, Mahmoud AM. 2023. A flavonoid-rich fraction of Euphorbia peplus attenuates hyperglycemia, insulin resistance, and oxidative stress in a type 2 diabetes rat model. *Frontiers in Pharmacology* **14**:1204641 DOI 10.3389/fphar.2023.1204641.

Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, Ogata H. 2020. KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36(7)**:2251–2252 DOI 10.1101/602110.

Arik SÖ, Pfister T. 2021. TabNet: attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence* **35(8)**:6679–6687 DOI 10.1609/aaai.v35i8.16826.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics* **25(1)**:25–29 DOI 10.1038/75556.

Borisov V, Leemann T, Seßler K, Haug J, Pawelczyk M, Kasneci G. 2022. Deep neural networks and tabular data: a survey. *IEEE Transactions on Neural Networks and Learning Systems* **35(6)**:7499–7519 DOI 10.1109/tnnls.2022.3229161.

Boscaro V, Holt CC, Van Steenkiste NWL, Herranz M, Irwin NAT, Àlvarez Campos P, Grzelak K, Holovachov O, Kerbl A, Mathur V, Okamoto N, Piercey RS, Worsaae K, Leander BS, Keeling PJ. 2022. Microbiomes of microscopic marine invertebrates do not reveal signatures of phylosymbiosis. *Nature Microbiology* **7(6)**:810–819 DOI 10.1038/s41564-022-01125-9.

Candela M, Biagi E, Soverini M, Consolandi C, Quercia S, Severgnini M, Peano C, Turroni S, Rampelli S, Pozzilli P, Pianesi M, Fallucca F, Brigidi P. 2016. Modulation of gut microbiota dysbioses in type 2 diabetic patients by macrobiotic Ma-Pi 2 diet. *British Journal of Nutrition* **116(1)**:80–93 DOI 10.1017/s0007114516001045.

Chari S, Acharya P, Gruen DM, Zhang O, Eyigoz EK, Ghalwash M, Seneviratne O, Saiz FS, Meyer P, Chakraborty P, McGuinness DL. 2023. Informing clinical assessment by contextualizing post-hoc explanations of risk prediction models in type-2 diabetes. *Artificial Intelligence in Medicine* **137(3)**:102498 DOI 10.1016/j.artmed.2023.102498.

Chen T, Guestrin C. 2016. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM, 785–794 DOI 10.1145/2939672.2939785.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34(17)**:i884–i890 DOI 10.1093/bioinformatics/bty560.

Chen X, Zhu Z, Zhang W, Wang Y, Wang F, Yang J, Wong K-C. 2022. Human disease prediction from microbiome data by multiple feature fusion and deep learning. *iScience* **25(4)**:104081 DOI 10.1016/j.isci.2022.104081.

Dai D, Zhu J, Sun C, Li M, Liu J, Wu S, Ning K, He L-J, Zhao X-M, Chen W-H. 2022. GMrepo v2: a curated human gut microbiome database with special focus on disease markers and cross-dataset comparison. *Nucleic Acids Research* **50(D1)**:D777–D784 DOI 10.1093/nar/gkab1019.

Dargan S, Kumar M, Ayyagari MR, Kumar G. 2020. A survey of deep learning and its applications: a new paradigm to machine learning. *Archives of Computational Methods in Engineering* **27(4)**:1071–1092 DOI 10.1007/s11831-019-09344-w.

Dash NR, Al Bataineh MT, Alili R, Al Safar H, Alkhayyal N, Prifti E, Zucker J-D, Belda E, Clément K. 2023. Functional alterations and predictive capacity of gut microbiome in type 2 diabetes. *Scientific Reports* **13(1)**:22386 DOI 10.1038/s41598-023-49679-w.

Dorogush AV, Ershov V, Gulin A. 2018. CatBoost: gradient boosting with categorical features support. ArXiv DOI 10.48550/arXiv.1810.11363.

Dwivedi K, Rajpal A, Rajpal S, Kumar V, Agarwal M, Kumar N. 2024. Enlightening the path to NSCLC biomarkers: utilizing the power of XAI-guided deep learning. *Computer Methods and Programs in Biomedicine* **243(5)**:107864 DOI 10.1016/j.cmpb.2023.107864.

Fan Y, Pedersen O. 2021. Gut microbiota in human metabolic health and disease. *Nature Reviews Microbiology* **19(1)**:55–71 DOI 10.1038/s41579-020-0433-9.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research* **39(suppl_2)**:W29–W37 DOI 10.1093/nar/gkr367.

Gao W, Lin W, Li Q, Chen W, Yin W, Zhu X, Gao S, Liu L, Li W, Wu D, Zhang G, Zhu R, Jiao N. 2024. Identification and validation of microbial biomarkers from cross-cohort datasets using xMarkerFinder. *Nature Protocols* **19(9)**:2803–2830 DOI 10.1038/s41596-024-00999-9.

Gao J, Yang T, Song B, Ma X, Ma Y, Lin X, Wang H. 2023. Abnormal tryptophan catabolism in diabetes mellitus and its complications: opportunities and challenges. *Biomedicine & Pharmacotherapy* **166**:115395 DOI 10.1016/j.biopha.2023.115395.

Gonzalez-Rellan MJ, Fernández U, Parracho T, Novoa E, Fondevila MF, da Silva Lima N, Ramos L, Rodríguez A, Serrano-Maciá M, Perez-Mejias G, Chantada-Vazquez P, Riobello C, Veyrat-Durebex C, Tovar S, Coppari R, Woodhoo A, Schwaninger M, Prevot V, Delgado TC, Lopez M, Diaz-Quintana A, Dieguez C, Guallar D, Frühbeck G, Diaz-Moreno I, Bravo SB, Martinez-Chantar ML, Nogueiras R. 2023. Neddylation of phosphoenolpyruvate carboxykinase 1 controls glucose metabolism. *Cell Metabolism* **35(9)**:1630–1645 DOI 10.1016/j.cmet.2023.07.003.

Gorishniy Y, Rubachev I, Khrulkov V, Babenko A. 2021. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems* **34**:18932–18943.

Gou W, Ling C-W, He Y, Jiang Z, Fu Y, Xu F, Miao Z, Sun T-Y, Lin J-S, Zhu H-L, Zhou H, Chen Y-M, Zheng J-S. 2021. Interpretable machine learning framework reveals robust gut microbiome features associated with type 2 diabetes. *Diabetes Care* **44(2)**:358–366 DOI 10.2337/dc20-1536.

Gurung M, Li Z, You H, Rodrigues R, Jump DB, Morgun A, Shulzhenko N. 2020. Role of gut microbiota in type 2 diabetes pathophysiology. *EBioMedicine* **51(10)**:102590 DOI 10.1016/j.ebiom.2019.11.051.

Han T, Srinivas S, Lakkaraju H. 2022. Which explanation should I choose? A function approximation perspective to characterizing post hoc explanations. *Advances in Neural Information Processing Systems* **35**:5256–5268.

Hancock JT, Khoshgoftaar TM. 2020. CatBoost for big data: an interdisciplinary review. *Journal of Big Data* **7(1)**:94 DOI 10.1186/s40537-020-00369-8.

Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. 1998. Support vector machines. *IEEE Intelligent Systems and their Applications* **13(4)**:18–28 DOI 10.1109/5254.708428.

Hugenholtz P, Tyson GW. 2008. Metagenomics. *Nature* **455(7212)**:481–483 DOI 10.1038/455481a.

Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11(1)**:673 DOI 10.1186/1471-2105-11-119.

Jang C, Oh SF, Wada S, Rowe GC, Liu L, Chan MC, Rhee J, Hoshino A, Kim B, Ibrahim A, Baca LG, Kim E, Ghosh CC, Parikh SM, Jiang A, Chu Q, Forman DE, Lecker SH, Krishnaiah S, Rabinowitz JD, Weljie AM, Baur JA, Kasper DL, Arany Z. 2016. A branched-chain amino acid metabolite drives vascular fatty acid transport and causes insulin resistance. *Nature Medicine* **22(4)**:421–426 DOI 10.1038/nm.4057.

Jensen LJ, Julien P, Kuhn M, von Mering C, Muller J, Doerks T, Bork P. 2007. eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Research* **36(suppl_1)**:D250–D254 DOI 10.1093/nar/gkm796.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **28(1)**:27–30 DOI 10.1093/nar/28.1.27.

Karlsson FH, Tremaroli V, Nookaew I, Bergström G, Behre CJ, Fagerberg B, Nielsen J, Bäckhed F. 2013. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498(7452)**:99–103 DOI 10.1038/nature12198.

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y. 2017. LightGBM: a highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems* **30**:3149–3157.

Krishna S, Han T, Gu A, Pombra J, Jabbari S, Wu S, Lakkaraju H. 2022. The disagreement problem in explainable machine learning: a practitioner's perspective. ArXiv DOI 10.48550/arXiv.2202.01602.

Lam TK, Gutierrez-Juarez R, Pocai A, Rossetti L. 2005. Regulation of blood glucose by hypothalamic pyruvate metabolism. *Science* **309(5736)**:943–947 DOI 10.1126/science.1112085.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9(4)**:357–359 DOI 10.1038/nmeth.1923.

Li Z. 2022. Extracting spatial effects from machine learning model using local interpretation method: an example of SHAP and XGBoost. *Computers, Environment and Urban Systems* **96(7)**:101845 DOI 10.1016/j.compenvurbsys.2022.101845.

Li Z, Liu F, Yang W, Peng S, Zhou J. 2021. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems* **33(12)**:6999–7019 DOI 10.1109/tnnls.2021.3084827.

Li M, Liu J, Zhu J, Wang H, Sun C, Gao NL, Zhao X-M, Chen W-H. 2023. Performance of gut microbiome as an independent diagnostic tool for 20 diseases: cross-cohort validation of machine-learning classifiers. *Gut Microbes* **15(1)**:2205386 DOI 10.1080/19490976.2023.2205386.

Li D, Luo R, Liu C-M, Leung C-M, Ting H-F, Sadakane K, Yamashita H, Lam T-W. 2016. MEGAHIT v1.0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102(6016)**:3–11 DOI 10.1016/j.ymeth.2016.02.020.

Li P, Zhu D. 2024. Clinical investigation of glucokinase activators for the restoration of glucose homeostasis in diabetes. *Journal of Diabetes* **16(5)**:e13544 DOI 10.1111/1753-0407.13544.

Liang P, Yang J, Wang W, Yuan G, Han M, Zhang Q, Li Z. 2023. Deep learning identifies intelligible predictors of poor prognosis in chronic kidney disease. *IEEE Journal of Biomedical and Health Informatics* **27(7)**:3677–3685 DOI 10.21203/rs.3.rs-1577772/v1.

Lind MV, Lauritzen L, Kristensen M, Ross AB, Eriksen JN. 2019. Effect of folate supplementation on insulin sensitivity and type 2 diabetes: a meta-analysis of randomized controlled trials. *The American Journal of Clinical Nutrition* **109(1)**:29–42 DOI 10.1093/ajcn/nqy234.

Liu Y, Liu Z, Luo X, Zhao H. 2022. Diagnosis of Parkinson's disease based on SHAP value feature selection. *Biocybernetics and Biomedical Engineering* **42(3)**:856–869 DOI 10.1016/j.bbe.2022.06.007.

Liu Y, Zhang Y, Imoto S. 2021. Discovering microbe functionality in human disease with a gene-ontology-aware model. In: *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Piscataway: IEEE, 1873–1880.

Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* **2(1)**:56–67 DOI 10.1038/s42256-019-0138-9.

Lundberg SM, Lee S-I. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* **30**:4768–4777.

Margeloiu A, Simidjievski N, Lio P, Jamnik M. 2023. Weight predictor network with feature selection for small sample tabular biomedical data. *Proceedings of the AAAI Conference on Artificial Intelligence* **37(8)**:9081–9089 DOI 10.1609/aaai.v37i8.26090.

Massart J, Zierath JR. 2019. Role of diacylglycerol kinases in glucose and energy homeostasis. *Trends in Endocrinology & Metabolism* **30(9)**:603–617 DOI 10.1016/j.tem.2019.06.003.

Menni C, Fauman E, Erte I, Perry JR, Kastenmüller G, Shin S-Y, Petersen A-K, Hyde C, Psatha M, Ward KJ, Yuan W, Milburn M, Palmer CN, Frayling TM, Trimmer J, Bell JT, Gieger C, Mohney RP, Brosnan MJ, Suhre K, Soranzo N, Spector TD. 2013. Biomarkers for type 2 diabetes and impaired fasting glucose using a nontargeted metabolomics approach. *Diabetes* **62(12)**:4270–4276 DOI 10.2337/db13-0570.

Pasolli E, Truong DT, Malik F, Waldron L, Segata N. 2016. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLOS Computational Biology* **12(7)**: e1004977 DOI 10.1371/journal.pcbi.1004977.

Pinaya WHL, Vieira S, Garcia-Dias R, Mechelli A. 2020. Autoencoders. In: *Machine Learning*. Amsterdam: Elsevier, 193–208.

Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D, Peng Y, Zhang D, Jie Z, Wu W, Qin Y, Xue W, Li J, Han L, Lu D, Wu P, Dai Y, Sun X, Li Z, Tang A, Zhong S, Li X, Chen W, Xu R, Wang M, Feng Q, Gong M, Yu J, Zhang Y, Zhang M, Hansen T, Sanchez G, Raes J, Falony G, Okuda S, Almeida M, LeChatelier E, Renault P, Pons N, Batto J-M, Zhang Z, Chen H, Yang R, Zheng W, Li S, Yang H, Wang J, Ehrlich SD, Nielsen R, Pedersen O, Kristiansen K, Wang J. 2012. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490(7418)**:55–60 DOI 10.1038/nature11450.

Ribeiro MT, Singh S, Guestrin C. 2016. "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*. New York: ACM, 1135–1144 DOI 10.1145/2939672.2939778.

Roy S, Laberge G, Roy B, Khomh F, Nikanjam A, Mondal S. 2022. Why don't XAI techniques agree? Characterizing the disagreements between post-hoc explanations of defect predictions. In: *2022 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. Piscataway: IEEE, 444–448.

Ruiz-Canela M, Guasch-Ferré M, Toledo E, Clish CB, Razquin C, Liang L, Wang DD, Corella D, Estruch R, Hernáez A, Yu E, Gómez-Gracia E, Arós F, Romaguera D, Dennis C, Ros E, Lapetra J, Serra-Majem L, Papandreou C, Portolés O, Fitó M, Salas-Salvadó J, Hu FB, Martínez-González MA. 2018. Plasma branched chain/aromatic amino acids, enriched Mediterranean diet and risk of type 2 diabetes: case-cohort study within the predimed trial. *Diabetologia* **61(7)**:1560–1571 DOI 10.1007/s00125-018-4611-5.

Sedighi M, Razavi S, Navab-Moghadam F, Khamseh ME, Alaei-Shahmiri F, Mehrtash A, Amirmozafari N. 2017. Comparison of gut microbiota in adult patients with type 2 diabetes and healthy individuals. *Microbial Pathogenesis* **111**:362–369 DOI 10.1016/j.micpath.2017.08.038.

Sereti I, Verburgh ML, Gifford J, Lo A, Boyd A, Verheij E, Verhoeven A, Wit FW, Schim van der Loeff MF, Giera M, Kootstra NA, Reiss P, Vujkovic-Cvijin I. 2023. Impaired gut microbiota-mediated short-chain fatty acid production precedes morbidity and mortality in people with HIV. *Cell Reports* **42(11)**:113336 DOI 10.1016/j.celrep.2023.113336.

Sharma A, Khade T, Satapathy SM. 2025. A cross dataset meta-model for hepatitis C detection using multi-dimensional pre-clustering. *Scientific Reports* **15(1)**:7278 DOI 10.1038/s41598-025-91298-0.

Sharma K, Saini N, Hasija Y. 2024. Identifying the mitochondrial metabolism network by integration of machine learning and explainable artificial intelligence in skeletal muscle in type 2 diabetes. *Mitochondrion* **74(2)**:101821 DOI 10.1016/j.mito.2023.11.004.

Shen Y, Zhu J, Deng Z, Lu W, Wang H. 2022. EnsDeepDP: an ensemble deep learning approach for disease prediction through metagenomics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **20(2)**:986–998 DOI 10.1109/tcbb.2022.3201295.

Shrikumar A, Greenside P, Kundaje A. 2017. Learning important features through propagating activation differences. In: *International Conference on Machine Learning*. PMLR, 3145–3153.

Simonyan K, Vedaldi A, Zisserman A. 2013. Deep inside convolutional networks: visualising image classification models and saliency maps. ArXiv DOI 10.48550/arXiv.1312.6034.

Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. 2017. SmoothGrad: removing noise by adding noise. ArXiv DOI 10.48550/arXiv.1706.03825.

Sun H, Saeedi P, Karuranga S, Pinkepank M, Ogurtsova K, Duncan BB, Stein C, Basit A, Chan JC, Mbanya JC, Pavkov ME, Ramachandaran A, Wild SH, James S, Herman WH, Zhang P, Bommer C, Kuo S, Boyko EJ, Magliano DJ. 2022. IDF Diabetes Atlas: global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Research and Clinical Practice* **183(10267)**:109119 DOI 10.1016/j.diabres.2021.109119.

Sundararajan M, Taly A, Yan Q. 2017. Axiomatic attribution for deep networks. In: *International Conference on Machine Learning*. PMLR, 3319–3328.

Tuppad A, Patil SD. 2024. An efficient classification framework for type 2 diabetes incorporating feature interactions. *Expert Systems with Applications* **239(4)**:122138 DOI 10.1016/j.eswa.2023.122138.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* **30**:6000–6010.

Wu H, Tremaroli V, Schmidt C, Lundqvist A, Olsson LM, Krämer M, Gummesson A, Perkins R, Bergström G, Bäckhed F. 2020. The gut microbiota in prediabetes and diabetes: a population-based cross-sectional study. *Cell Metabolism* **32(3)**:379–390 DOI 10.1016/j.cmet.2020.06.011.

Xue C-X, Lin H, Zhu X-Y, Liu J, Zhang Y, Rowley G, Todd JD, Li M, Zhang X-H. 2021. DiTing: a pipeline to infer and compare biogeochemical pathways from metagenomic and metatranscriptomic data. *Frontiers in Microbiology* **12**:698286 DOI 10.3389/fmicb.2021.698286.

Yamada K, Mendoza J, Koutmos M. 2023. 5-formyltetrahydrofolate promotes conformational remodeling in a methylenetetrahydrofolate reductase active site and inhibits its activity. *Journal of Biological Chemistry* **299(2)**:102855 DOI 10.1016/j.jbc.2022.102855.

Yang X, Hu R, Wang Z, Hou Y, Song G. 2023. Associations between serum folate level and HOMA-IR in Chinese patients with type 2 diabetes mellitus. *Diabetes, Metabolic Syndrome and Obesity* **16**:1481–1491 DOI 10.2147/dmso.s409291.

Zhang Y-H, Guo W, Zeng T, Zhang S, Chen L, Gamarra M, Mansour RF, Huang T, Cai Y-D. 2021. Identification of microbiota biomarkers with orthologous gene annotation for type 2 diabetes. *Frontiers in Microbiology* **12**:711244 DOI 10.3389/fmicb.2021.711244.

Zhao Y, Shao J, Asmann YW. 2022. Assessment and optimization of explainable machine learning models applied to transcriptomic data. *Genomics, Proteomics & Bioinformatics* **20(5)**:899–911 DOI 10.1016/j.gpb.2022.07.003.

Zhao L, Zhang F, Ding X, Wu G, Lam YY, Wang X, Fu H, Xue X, Lu C, Ma J, Yu L, Xu C, Ren Z, Xu Y, Xu S, Shen H, Zhu X, Shi Y, Shen Q, Dong W, Liu R, Ling Y, Zeng Y, Wang X, Zhang Q, Wang J, Wang L, Wu Y, Zeng B, Wei H, Zhang M, Peng Y, Zhang C. 2018. Gut bacteria selectively promoted by dietary fibers alleviate type 2 diabetes. *Science* **359(6380)**:1151–1156 DOI 10.1126/science.aao5774.

Zheng Y, Ley SH, Hu FB. 2018. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. *Nature Reviews Endocrinology* **14(2)**:88–98 DOI 10.1038/nrendo.2017.151.

Zhou H, Wang X, Zhu R. 2022. Feature selection based on mutual information with correlation coefficient. *Applied Intelligence* **52**:1–18 DOI 10.1007/s10489-021-02524-x.