OXFORD

Briefings in Bioinformatics, 2025, **26(3)**, bbaf217
https://doi.org/10.1093/bib/bbaf217
Problem Solving Protocol

MSFT-transformer: a multistage fusion tabular transformer for disease prediction using metagenomic data

Ning Wang[1], Minghui Wu[1], Wenchao Gu[1], Chenglong Dai[1,*], Zongru Shao[2], K.P. Subbalakshmi[3]

[1]School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214121, Jiangsu, China
[2]Silicon Austria Labs, Linz, Austria
[3]Department of Electrical and Computer Engineering, Stevens Institute of Technology, Castle Point Terrace, Hoboken, NJ 07030, United States
*Corresponding author: School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214121, Jiangsu, China.
E-mail: chenglongdai@jiangnan.edu.cn

Abstract

More and more recent studies highlight the crucial role of the human microbiome in maintaining health, while modern advancements in metagenomic sequencing technologies have been accumulating data that are associated with human diseases. Although metagenomic data offer rich, multifaceted information, including taxonomic and functional abundance profiles, their full potential remains underutilized, as most approaches rely only on one type of information to discover and understand their related correlations with respect to disease occurrences. To address this limitation, we propose a multistage fusion tabular transformer architecture (MSFT-Transformer), aiming to effectively integrate various types of high-dimensional tabular information extracted from metagenomic data. Its multistage fusion strategy consists of three modules: a fusion-aware feature extraction module in the early stage to improve the extracted information from inputs, an alignment-enhanced fusion module in the mid stage to enforce the retainment of desired information in cross-modal learning, and an integrated feature decision layer in the late stage to incorporate desired cross-modal information. We conduct extensive experiments to evaluate the performance of MSFT-Transformer over state-of-the-art models on five standard datasets. Our results indicate that MSFT-Transformer provides stable performance gains with reduced computational costs. An ablation study illustrates the contributions of all three models compared with a reference multistage fusion transformer without these novel strategies. The result analysis implies the significant potential of the proposed model in future disease prediction with metagenomic data.

Keywords: disease prediction; human microbiome; multimodality; multistage fusion; tabular transformer

Introduction

The human gut microbiota is pivotal in influencing various diseases and maintaining health through its collective metabolic and immune interactions with the host [1]. Researchers have been capturing comprehensive snapshots of the gut microbial community using advanced high-throughput sequencing technologies, resulting in a substantial accumulation of relevant genomic data. These sequencing technologies have enabled a sophisticated understanding by profiling the taxonomic composition and functional potentials within the targeted microbial communities [2, 3]. Consequently, this has spurred a series of related studies and rapid advancements in understanding the association between these diseases and gut microbiota, particularly with the integration of artificial intelligence development and machine learning modeling.

Within this scope, the early detection of several diseases has been frequently evaluated in recent years due to their confirmed correlation with metagenomic information [4–6]. Some of these studies have investigated the modeling of taxonomic profiles, incorporating conventional machine learning algorithms and state-of-the-art (SOTA) deep neural networks. For example, DeepMicro [7] developed a deep learning framework that uses auto-encoder (AE) feature extraction strategies to transform high-dimensional microbiome profiles into low-dimensional representations, along with downstream classifiers for the detection of several diseases, including European women type 2 diabetes (EW-T2D) [8], liver cirrhosis (LC) [9], Chinese type 2 diabetes (C-T2D) [10], inflammatory bowel disease (IBD) [11], and obesity [12]. This approach achieved promising results, with observations indicating that the effectiveness of different types of AE varied depending on the complexity and characteristics of the data. These findings underscore the importance of selecting the appropriate architecture to achieve optimal performance. MegaR [13] was introduced as both a web application and an R Shiny package for interactive analysis and visualization, utilizing taxonomic profiles within this domain. EnsDeepDP [4] was also benchmarked on these datasets, specializing in feature selection. Another branch of works, including PopPhy-CNN [14],

Received: January 26, 2025. **Revised:** April 05, 2025. **Accepted:** April 21, 2025

Meta-Signer [15], TaxoNN [16], EPCNN [17], and MicroKPNN [18], aimed to enhance the utilization of taxonomic profiles by integrating knowledge from phylogenetic trees. Among them, PopPhy-CNN utilized convolutional neural networks (CNNs) to extract features from a flattened phylogenetic tree that integrates relative taxonomic abundance information. Meta-Signer converted the phylogenetic tree into embedding vectors before feeding them into CNNs. TaxoNN focused on extracting high-correlation information from an operational taxonomic unit table based on phylogenetic trees.

Separate from taxonomic profiles, several studies have explored the functional potentials of the gut microbiome, given their contextual information regarding disease pathways and metabolic insights. These studies primarily rely on functional annotations from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [19], Gene Ontology (GO) [20], and eggNOG [21]. For instance, Zhang *et al.* [22] annotated functional features using KEGG and eggNOG for type 2 diabetes (T2D). Liu *et al.* [23] annotated gene abundance with GO terms. Both studies focused on SOTA machine learning models due to the large feature dimension. In summary, the aforementioned studies have systematically explored the gut microbiome at both taxonomic and functional levels. However, these studies relied heavily on single-modality approaches, which may fail to capture a more comprehensive understanding of microbiome–disease associations.

Regarding the most recent advances in multimodal approaches within the scope of deep learning, transformer architectures [24] have demonstrated significant potential in multimodal contexts [25]. For example, the Multimodal Bottleneck Transformer (MBT) [26] introduced bottleneck tokens to facilitate the early- and middle-stage fusion of audio and visual signals in videos.

Given the distinct statistical characteristics of taxonomic profiles and functional features, they are considered different modalities of genomic data. Consequently, several studies have attempted to fuse useful information between them to achieve further performance improvements. An intuitive concatenation of the two feature sets with an artificial neural network (ANN) [27] has been investigated. Further improvement has been achieved by concatenating multimodal embeddings in MDL4Microbiome [28]. MVIB [29], a multimodal variational information bottleneck model, enhanced the joint learning of the representations across different modalities through a product-of-experts approach.

Unlike previous approaches, we investigate potential performance gains of different fusion strategies with metagenomic data without requiring domain-specific knowledge, focusing on extracted features from taxonomic profiles and functional annotations. Furthermore, we propose a novel multistage multimodal architecture, MSFT-Transformer, as shown in Fig. 1. In summary, our major contributions are as follows:

1. We introduce the MSFT-Transformer, which integrates taxonomic and functional information through a multistage fusion process, providing a more robust framework for disease prediction without requiring domain-specific knowledge.
2. The MSFT-Transformer employs a three-stage fusion process:
    - In the early-stage fusion, we propose a fusion-aware initialization strategy for optimizing the initialization of the bottleneck tokens.
    - In the middle-stage fusion, we introduce two modules: the *Intra-Modality Alignment Module* (IMAM), which focuses on learning intra-modality relationships, and the *Cross-Modality Fusion Module* (CMFM), designed to capture and integrate cross-modal interactions.
    - In the late-stage fusion, we integrate the modal-specific information with fully fused cross-modality information for the final decision-making.
3. Extensive experiments and comparisons with SOTA models on five real-world metagenomic datasets show that our approach outperforms others in terms of all three performance metrics often in our evaluation.

Materials and methods

This section presents the design of the Multi-Stage Fusion Tabular (MSFT-) Transformer following a brief description of data preprocessing in our pipeline.

Data preprocessing from raw metagenomic data

Raw metagenome sequencing data obtained in the aforementioned previous studies are used as inputs. Two types of microbiome features are extracted from the metagenomic data in five datasets [8–12]: (1) taxonomic features, represented by the species-level relative abundance and (2) functional features, represented by the relative abundances of each KEGG Orthology (KO). Note that KOs are a collection of manually defined ortholog groups of functionally equivalent gene sets across different organisms. We directly use the taxonomic features extracted from DeepMicro [7], where MetaPhlAn2 [30] was employed to extract the abundance profiles with the default parameters. Subsequently, MetAML [31] was utilized to preprocess these abundance profiles by specifically selecting desired species-level features while excluding sub-species-level features. Functional features are further extracted in parallel from raw metagenome sequencing data in three steps. (a) We employ the fastp (an ultra-fast FASTQ preprocessor with quality control and data-filtering features) [32] for quality control. (b) Then, human genomes are removed using Bowtie2 [33], where we use Genome Reference Consortium Human Build 38 (GRCh38) as the reference. (3) Finally, we extract KO relative abundances using the DiTing [34] tool on the remaining reads. We filter out the features with relative abundances below 0.05% across all samples [35]. This filtration process is applied independently to both taxonomic and functional features.

Architecture of MSFT-transformer

The architecture of MSFT-Transformer is illustrated in Fig. 2. The model integrates tabular taxonomic and functional features with a three-stage fusion process that enhances the fusion of two modalities in each stage. (a) The early-stage fusion employs a *Fusion-Aware Feature Extraction Module* (FAFEM) with a fusion-aware initialization strategy. (b) The middle-stage utilizes an *Alignment-Enhanced Fusion Module* (AEFM), which consists of two key submodules: an IMAM and a CMFM. (3) Finally, the late-stage benefits from an *IFDL* before the final classification. A comprehensive description of the design is presented below sequentially.

Early-stage: Fusion-Aware Feature Extraction Module

The FAFEM aims to extract rich features from different preprocessed modalities at an early stage and prepare them for subsequent fusion processes. Given that the preprocessed features

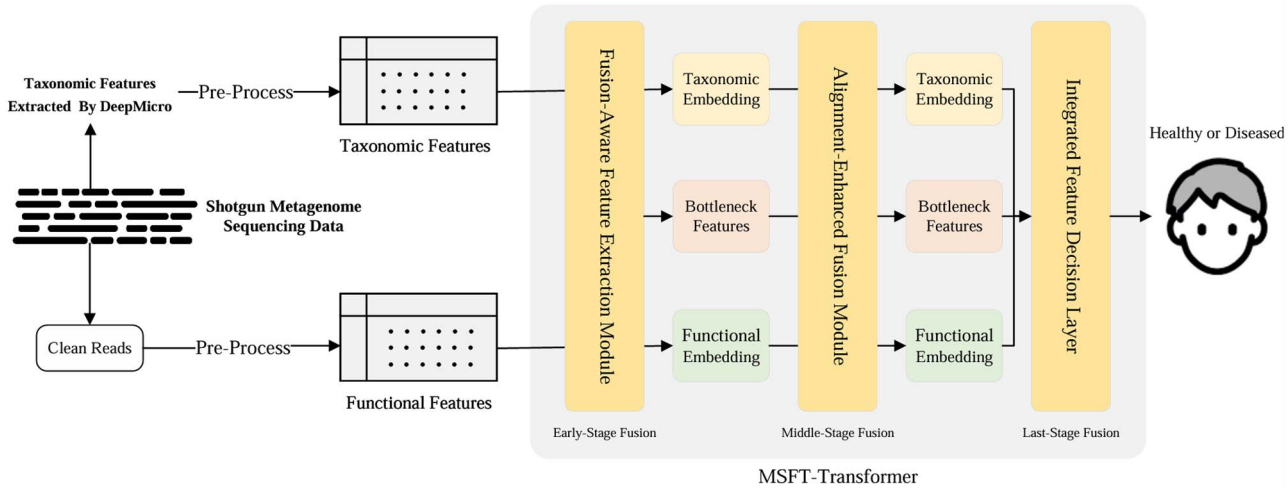

Figure 1. An overview of the proposed framework, illustrating the overall data processing and model architecture, which consists of multistage fusion: early-stage fusion—FAFEM; middle-stage fusion—AEFM; late-stage fusion—IFDL.

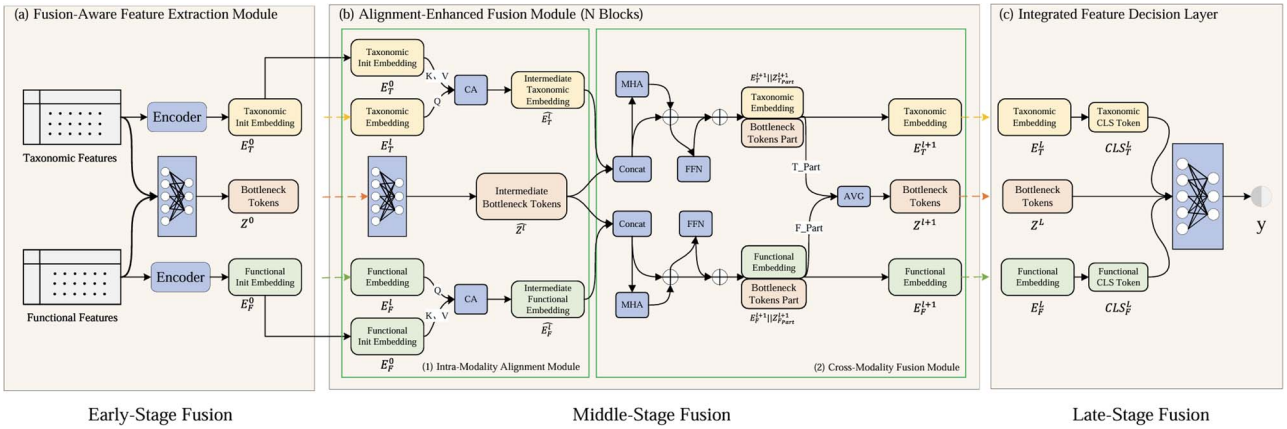

Figure 2. The overall architecture of MSFT-Transformer with fusion in three stages: (a) early-, (b) middle-, and (c) late-stage fusion, while (b) middle-stage fusion consists of two submodules (b).1. IMAM and (b).2. CMFM.

are tabular, we adapt the tabular inputs to the transformer architecture based on the adaptation strategy in a SOTA scheme, FT-Transformer [36], as it outperforms several counterpart models in several related tabular benchmarks. Note that the FT-Transformer is unimodal. Thus, we adapt each of the modalities and fuse them in FAFEM. The complete process is described as below. The encoder in FT-Transformer is adopted to transform the features into embeddings in two modalities: E_T for taxonomic features and E_F for functional annotations. We denote them as E_i, $i \in \{T, F\}$, and the encoder is formulated as follows:

$$\begin{aligned} E_{i,j_i} &= b_{i,j_i} + x_{i,j_i} \cdot W_{i,j_i} \\ E_i &= stack[E_{i,1}, E_{i,2}, \ldots E_{i,N_i}], \end{aligned} \tag{1}$$

where j_i is the input index of the features under modal i and $j_i \in \{1, 2, \ldots, N_i\}$. b_{i,j_i} and W_{i,j_i} are the jth feature bias and weight vector of modal i, respectively. And N_i is the dimension of modal i. Then, we append the classification tokens (CLS) [37] to both modalities to obtain the initial embeddings E^0:

$$E_i^0 = stack([CLS], E_i). \tag{2}$$

After deriving E_T^0 and E_F^0, it is intuitive to apply a naïve concatenation. However, it introduces a significant computational complexity [25] given the nontrivial embedding dimensions. Therefore, bottleneck tokens [26, 38, 39] are introduced in FAFEM to mitigate the quadratic complexity of pairwise attention. Meanwhile, bottleneck tokens serve the role of cross-modal information learning. As a result, bottleneck tokens function as a bridge for information integration, transmitting the condensed unimodal information to another modality.

Given that the bottleneck tokens encapsulate the essential information from both modalities and play a critical role during the fusion process, we propose a fusion-aware initialization strategy to initialize them to improve the learned cross-modal information and enhance their effectiveness. Such an enhancement of early-stage fusion is motivated by a psychiatric discovery: biological neural networks engage in multimodal fusion at the earliest layers of sensory processing pathways [40]. Meanwhile, early fusion in ANN can effectively capture cross-modal relationships from low-level features [41, 42]. Recognizing the critical contributions of early-stage cross-modal learning through improved weight initialization for optimal performance [43], we initialize the bottleneck tokens by employing early fusion on features from both input modalities. The initialization is formulated as follows:

$$Z^0 = MLP(stack[x_F, x_T]), \tag{3}$$

where x_F and x_T are the functional and taxonomic features, respectively. And Z^0 denotes the bottleneck tokens initialized from the original features. Thus, FAFEM provides three outputs: E_T^0, Z^0, and E_F^0.

Middle-stage: Alignment-Enhanced Fusion Module

The AEFM consists of an IMAM and CMFM, as shown in Fig. 2(b), to improve the learning in both intra- and cross-modal directions. Note that AEFM can be stacked into L blocks, enabling deeper integration of information and enhancing the effectiveness of tokens and embeddings. The design of IMAM and CMFM is described as below.

IMAM The objective of IMAM is to elevate intra-modality information extraction in AEFM through improved alignment with the original modality. It is necessary for multimodal fusion to preserve intra-modality information besides exploiting cross-modality information [44]. However, over-fusion of cross-modality embeddings may result in the loss of the information from its original modality. We specifically design IMAM to overcome the potential loss, as shown in Fig. 2(b).1. To achieve this, we input the initial embeddings (E_F^0 and E_T^0) into all AEFM blocks [45] and align the unimodal embeddings with their original values through *cross-attention* mechanism [46]. We denote E_i^l as the embedding of the lth block given that $i \in \{T, F\}$ and $l \in \{0, 1, 2, ..., L-1\}$. Then, E_i^0 serves as both the key and value in cross-attention, providing a stable reference point for alignment, while E_i^l, updated through the IMAM in each of the stacked AEFM blocks, functions as the query to obtain the intermediate intra-modal embeddings $\widehat{E_i^l}$. Further, the intermediate bottleneck tokens $\widehat{Z^l}$ are updated by a multilayer perceptron (MLP) for capturing and refining the shared latent space across modalities. $\widehat{E_i^l}$ and $\widehat{Z^l}$ are formulated as follows:

$$CrossAttention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{D}}\right)V \tag{4}$$

$$\widehat{E_i^l} = CrossAttention(E_i^l, E_i^0, E_i^0) \tag{5}$$

$$\widehat{Z^l} = MLP(Z^l), \tag{6}$$

where D is the dimension of the key embeddings (E_i^0). As a result, IMAM passes $\widehat{E_T^l}$, $\widehat{Z^l}$, and $\widehat{E_F^l}$ to the following CMFM while strengthening all three through their enhanced intra-modal alignment, thus, mitigating potential information loss.

CMFM Following the intra-modal alignment, CMFM, as shown in Fig. 2(b).2, aims to achieve effective cross-modal fusion with $\widehat{E_T^l}$, $\widehat{Z^l}$, and $\widehat{E_F^l}$, considering the complementarity and redundancy of multimodal information [42]. CMFM computes the outputs of the lth AEFM block: E_T^{l+1}, Z^{l+1}, and E_F^{l+1}, as shown below:

$$AttentionResidual(x) = x + MHA(LN(x)) \tag{7}$$

$$TransformerBlock(x) = AttentionResidual(x) \tag{8}$$

$$+ FFN(LN(AttentionResidual(x)) \tag{9}$$

$$[E_F^{l+1}||Z_{F_{Part}}^{l+1}] = TransformerBlock[\widehat{E_F^l}||\widehat{Z^l}] \tag{10}$$

$$[E_T^{l+1}||Z_{T_{Part}}^{l+1}] = TransformerBlock[\widehat{E_T^l}||\widehat{Z^l}], \tag{11}$$

Note that *MHA* denotes the *MultiHeadAttenion* mechanism, *LN* Layer Normalization, and *FFN* Feed Forward Network. $Z_{F_{Part}}^{l+1}$ and $Z_{T_{Part}}^{l+1}$ are the bottleneck fusion tokens that capture the compressed feature information from the functional and taxonomic modalities, respectively. The output bottleneck tokens are calculated by $Z_{F_{Part}}^{l+1}$ and $Z_{T_{Part}}^{l+1}$:

$$Z^{l+1} = Average\left(Z_{F_{Part}}^{l+1}, Z_{T_{Part}}^{l+1}\right) \tag{12}$$

Thus, CMFM achieves two objectives: (a) the unimodal features are learning the cross-modal information through the bottleneck tokens; (b) the bottleneck tokens also preserve modality-specific information, effectively balancing cross-modal integration besides the retention of modality-specific data.

Late-stage: Integrated Feature Decision Layer

We propose enhanced late-stage fusion with bottleneck tokens in the classification layer since they are dimension-reduced for the condensation and integration of essential information from both modalities. As a compact representation and a cross-modal balancer, their potential for decision-making cannot be dismissed. Therefore, we incorporate an IFDL to utilize their information. Thus, we extract the unimodal classification tokens, CLS_T^L and CLS_F^L, and then concatenate them with the last-layer bottleneck tokens Z^L. The joint representation is subsequently passed to an MLP [47] for the final decision-making:

$$\hat{y} = \text{MLP}(\left[CLS_T^L||Z^L||CLS_F^L\right], \tag{13}$$

where $\hat{y}$ denotes the final prediction.

Experiments and results

This section presents the experiments and their results, including a brief description of the datasets, the comparison baselines, the experimental setup, and a detailed evaluation along with its discussion. A performance analysis and an ablation study are also included at the end.

Datasets

We benchmark the proposed model on five standard metagenomic datasets for disease prediction, including EW-T2D [8], LC [9], C-T2D [10], IBD [11], and Obesity [12]. Each dataset contains samples from both patients and healthy controls, which are statistically shown in Table 1. There are 1035 human gut metagenomic samples in total from these five datasets. Note that three out of five datasets are considered balanced, while the other two (IBD and Obesity) are relatively imbalanced according to the ratio of patients (Positive, P) over healthy controls (Negative, N). In addition, the dimensions of both taxonomic and functional features, including the original dimensions and those after feature filtering at 0.05%, are shown in Table 2.

Baselines

To conduct a direct comparison with SOTA methods, we compare MSFT-Transformer with both SOTA uni- and multi-modal models. The former includes RF [48], SVM [49], MLP [47], FT-Transformer [36], DeepMicro [7], EnsDeepDP [4], and several unimodal taxonomic models with external knowledge. The latter includes several single-stage fusion methods, such as FT-Concat and FT-Vote (naïve multimodal adaptations of FT-Transformer for early- and late-fusion respectively), MDL4Microbiome [28], and MVIB [29]. We also include T-MBT (modified from MBT [26] for the adaptation of tabular inputs) as a representative of multistage fusion methods.

Table 1. Benchmark datasets

Dataset	$N_{samples}$	$N_{patient}$	$N_{healthy}$	P/N Ratio	Balance (Y/N)
EW-T2D	96	53	43	1.13: 1	Y
LC	232	118	114	1: 1	Y
C-T2D	344	170	174	1: 1	Y
IBD	110	25	85	1: 3.4	N
Obesity	253	164	89	1.84: 1	N

Table 2. Comparison of microbiome feature dimensions before and after filtering at a 0.05% abundance threshold

Feature	Taxonomic		Functional	
	Original	Filtered (0.05%)	Original	Filtered (0.05%)
EW-T2D	381	269	7155	1315
LC	542	377	7465	1980
C-T2D	572	372	7434	2025
IBD	443	278	6282	1927
Obesity	465	282	7606	1907

Experimental setup

Our experiments are conducted using Pytorch 1.12.1 on two NVIDIA A40 GPUs and the Scikit-learn Python library [50]. To increase the reproducibility of the results, we conducted five trials, each utilizing a different random seed for the splitting process: 64% training, 16% validation, and 20% testing. The averaged evaluation results over all five trials are reported in the end (All relevant hyperparameter settings and training configurations can be found in our GitHub repository).

Evaluation and discussion

Our evaluation is based on the conventional metrics on the five standard datasets: area under curve (abbreviated as AUC), recall, and precision. A performance comparison of all benchmarked models is shown in Table 3. Meanwhile, an analysis of model stability concerning the receiver operating characteristic (ROC) curve is shown in Fig. 3. We compare the proposed MSFT-Transformer to other baseline models from two aspects: (1) multimodal versus unimodal and (2) single- and multistage fusion versus. MSFT-Transformer.

Multimodal versus unimodal models

The comparison of multimodal and unimodal models focuses on blocks A and C in Table 3, where the former utilizes the taxonomic features only, while the latter includes function information as well. Models using taxonomic features but assisted by external knowledge are excluded from this comparison as they are visibly benefiting from the additional resources in block B. Several observations can be made by comparing AUC (A), precision (P), and recall (R) across blocks A and C. (1) Multimodal methods do not necessarily outperform unimodal methods. For example, RF (unimodal) outperforms MDL4Microbiome (multimodal) w.r.t. AUC for all five datasets (78.2%>76.8%, 95.1%>92.5%, 75.1%>74.7%, 92.6%>88.0%, and 65.3%>54.1%). It implies that the simplicity of unimodal models leads to an easier learning process, while the complexity of multimodal methods introduces learning barriers. Excluding the proposed MSFT-Transformer, the previously reported results of EnsDeepDP achieved the best performance, under the condition that their train-test partition was different from ours. (2) Boosted performance of MSFT-Transformer is apparent. From the nine models benchmarked with our framework, MSFT-Transformer achieved the best AUC in four out of the five datasets, in contrast to MDL4Microbiome compared with RF, SVM, MLP, and FT-Transformer. It shows that the proposed MSFT-Transformer does effectively learn from both modalities.

When including block B in the comparison, it shows that external knowledge significantly contributes to the disease prediction in two datasets, while it is not so visible in the other three. However, these external resources often require extra effort from domain experts and significant data curation. It implies that relevant external knowledge may enhance performance, but there are also failing times. In contrast, MSFT-Transformer outperforms a majority of these methods in blocks A, B, and C (beats 12/12, 12/15, 12/13, 13/13, and 12/14 methods in five datasets, respectively), requiring no extra knowledge. Therefore, MSFT-Transformer provides a more cost-effective solution with stable performance gains.

In summary, it indicates that multimodal fusion introduces computational complexity and learning barriers, while an enhanced fusion approach can achieve stable performance gains across datasets and is more cost-effective toward external resource requirements.

Single-stage and multistage fusion versus MSFT

This comparison reveals the effectiveness of different fusion strategies by comparing benchmarks in block C of Table 3. There are several observations. (1) *Transformer versus ANN.* Single-stage transformer architectures (FT-Concat and FT-Vote) outperform MDL4Microbiome for four out of five datasets (EW-T2D, C-T2D, IBD, and Obesity), given the complete alignment in experimental setup. (2) *Early versus late fusion.* Early fusion with FT-Concat shows trivial differences compared with late fusion with FT-Vote. Only for IBD, there is a 2.6% difference w.r.t. AUC (compared with differences of 0.4%, 0.3%, 0.9%, and 0.5% for the other four datasets, respectively). (3) *Single- versus Multistage Fusion.* Compared with single-stage fusion models, FT-Concat and FT-Vote, multistage fusion T-MBT achieves improved performance for three out of five datasets, with one tied (LC) and the other falling behind (C-T2D). It shows that multistage fusion does have

Table 3. Performance evaluation with AUC (A), precision (P), and recall (R) (with the standard mean error in brackets)

Comparison	Method		EW-T2D	LC	C-T2D	IBD	Obesity
A. Taxonomic features	RF	A	78.2%(3.2%)	95.1%(1.5%)	75.1%(2.6%)	92.6%(2.2%)	65.3%(4.4%)
		P	71.3%(2.3%)	**92.6%(1.7%)**	66.9%(1.9%)	70.0%(20.0%)	64.1%(0.4%)
		R	80.0%(3.4%)	82.5%(3.6%)	62.9%(3.9%)	20.0%(6.3%)	**97.6%(1.5%)**
	SVM	A	75.5%(3.1%)	89.4%(2.2%)	71.6%(2.4%)	92.0%(2.6%)	64.2%(3.8%)
		P	75.2%(3.6%)	81.9%(3.3%)	65.7%(3.5%)	66.7%(18.3%)	70.9%(2.5%)
		R	69.1%(6.2%)	75.0%(4.2%)	61.8%(4.7%)	36.0%(9.8%)	77.0%(3.1%)
	MLP	A	74.1%(5.1%)	90.0%(1.5%)	72.3%(2.2%)	86.4%(5.0%)	52.3%(3.0%)
		P	**86.4%(6.9%)**	79.3%(2.1%)	67.3%(2.0%)	86.3%(2.3%)	67.2%(1.5%)
		R	69.1%(6.8%)	86.1%(3.2%)	64.1%(5.8%)	**94.1%(2.6%)**	70.3%(1.5%)
	FT-Transformer	A	78.4%(3.6%)	90.4%(2.4%)	73.2%(3.0%)	93.6%(2.0%)	66.1%(1.7%)
		P	71.4%(4.7%)	84.8%(6.2%)	66.8%(3.2%)	70.1%(7.7%)	**75.9%(2.5%)**
		R	69.1%(6.8%)	85.8%(3.4%)	64.1%(3.3%)	72.0%(4.9%)	67.3%(6.9%)
	DeepMicro [7]	A	82.9%(3.9%)	88.8%(1.1%)	72.5%(2.5%)	87.3%(3.0%)	67.4%(3.4%)
		P	–	–	–	–	–
		R	–	–	–	–	–
	EnsDeepDP [4]	A	86.7%(N/A)	94.3%(N/A)	**77.6%(N/A)**	95.8%(N/A)	72.3%(N/A)
		P	–	–	–	–	–
		R	–	–	–	–	–
B. Taxonomic features with external knowledge	Meta-Signer [15]	A	–	90.5%(5.0%)	–	79.4(15.9%)	60.0%(13.5%)
		P	–	–	–	–	–
		R	–	–	–	–	–
	TaxoNN [16]	A	–	91.1%(N/A)	73.3%(N/A)	–	–
		P	–	–	–	–	–
		R	–	–	–	–	–
	PopPhy-CNN [14]	A	–	90.1%(N/A)	–	–	58.9%(N/A)
		P	–	–	–	–	–
		R	–	–	–	–	–
	MicroKPNN [18]	A	85.8(6.7%)	**96.9%(0.9%)**	75.5%(3.2%)	95.4%(3.7%)	**72.8%(4.8%)**
		P	–	–	–	–	–
		R	–	–	–	–	–
C. Taxonomic features and functional features	MDL4Microbiome	A	76.8%(5.7%)	92.5%(1.3%)	74.7%(1.3%)	88.0%(0.6%)	54.1%(2.0%)
		P	76.3%(7.5%)	78.1%(1.4%)	67.6%(1.9%)	86.0%(1.7%)	69.3%(1.9%)
		R	69.1%(4.6%)	89.1%(5.4%)	66.5%(4.2%)	92.9%(2.9%)	72.1%(3.2%)
	MVIB [29]	A	85.9%(2.3%)	92.5%(0.5%)	75.8%(1.2%)	93.6%(1.4%)	66.6%(2.7%)
		P	–	–	–	–	–
		R	–	–	–	–	–
	FT-Concat	A	82.6%(3.1%)	91.7%(1.7%)	77.5%(1.9%)	92.0%(2.4%)	61.5%(4.2%)
		P	77.8%(4.0%)	89.4%(3.2%)	68.9%(2.2%)	79.3%(9.7%)	75.1%(3.3%)
		R	80.0%(7.3%)	76.7%(4.5%)	68.2%(6.1%)	64.0%(7.5%)	57.6%(12.4%)
	FT-Vote	A	82.2%(2.1%)	91.4%(1.4%)	76.6%(1.6%)	94.6%(1.7%)	61.0%(3.3%)
		P	70.5%(5.4%)	84.4%(4.0%)	68.3%(1.5%)	77.7%(6.1%)	70.4%(2.1%)
		R	80.0%(6.0%)	87.5%(2.9%)	63.5%(4.3%)	56.0%(11.7%)	63.0%(11.9%)
	T-MBT	A	83.2%(4.2%)	91.7%(2.3%)	76.6%(1.6%)	95.5%(2.1%)	62.2%(3.3%)
		P	80.0%(2.8%)	87.1%(4.2%)	69.7%(2.8%)	81.7%(7.6%)	73.1%(2.6%)
		R	67.3%(6.2%)	80.8%(6.0%)	**70.0%(1.7%)**	68.0%(8.0%)	60.6%(7.3%)
	MSFT-Transformer	A	**87.5%(4.1%)**	94.1%(1.2%)	**77.6%(1.6%)**	**97.4%(1.8%)**	67.7%(2.9%)
		P	80.5%(7.0%)	88.7%(3.9%)	**70.8%(2.8%)**	**100.0%(0.0%)**	72.7%(2.9%)
		R	**83.6%(6.0%)**	**89.2%(3.4%)**	65.3%(2.5%)	72.0%(13.6%)	69.7%(10.5%)

The reported results of cited methods are based on previous works that use the same datasets. There are a total of nine methods (uncited) evaluated with our framework for five trails, including MSFT-Transformer. N/A indicates that the corresponding study did not provide the standard mean error. The best performance of each dataset is highlighted in bold while the second-high underlined.

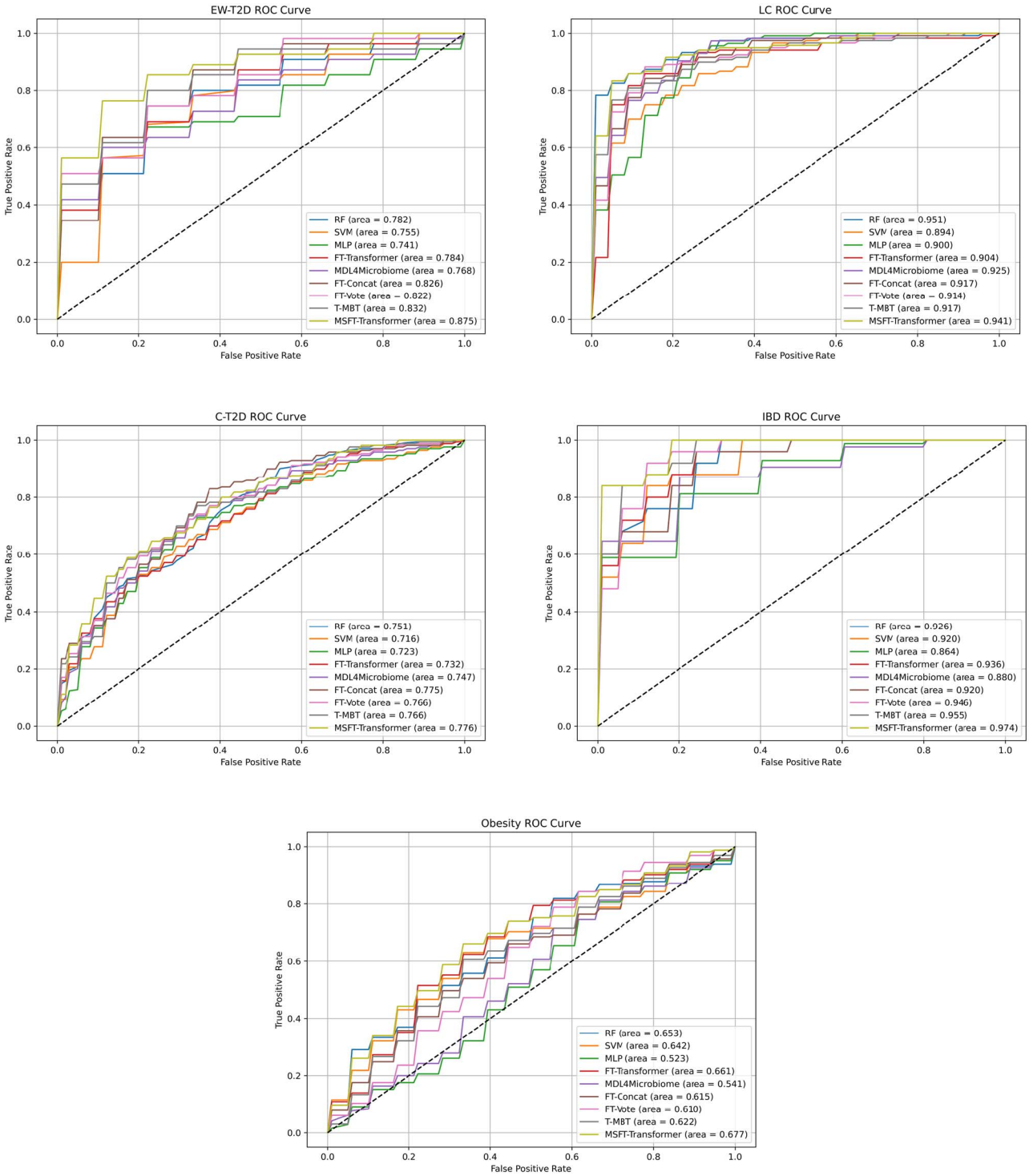

Figure 3. ROC curves and AUCs for MSFT-Transformer and baseline models we experimented with.

the potential for better cross-modal learning, while it does not always excel due to the complexity of data. (4) *T-MBT versus MSFT-Transformer.* There is an apparent performance increase of MSFT-Transformer compared with T-MBT although both of them are multistage fusion strategies. This indicates that MSFT-Transformer achieves enhanced cross-modal learning through multiple stages.

In summary, multistage fusion has apparent potential and advantages in cross-modal learning compared with single-stage fusion. Its sophisticated design and computational complexity may hinder its performance on specific datasets. There is clear evidence that the proposed MSFT-Transformer strengthens multistage learning with its novel strategies in different stages.

Robustness evaluation

To evaluate the robustness of MSFT-Transformer, we conducted robustness tests on five datasets against other multimodal models, as shown in Table 4. Specifically, we performed experiments under three levels of Gaussian noise: 0.05, 0.1, and 0.3. By introducing controlled Gaussian noise at varying intensities, we

Table 4. Robustness evaluation of different models under varying levels of Gaussian noise across five datasets

Disease	noise level	FT-Concat	FT-Vote	T-MBT	MSFT-Transformer
EW-T2D	0.3	80.2%(3.9%)	82.6%(4.4%)	79.2%(3.0%)	**84.7%(2.5%)**
	0.1	80.4%(2.7%)	80.4%(3.4%)	79.2%(3.8%)	**85.1%(4.0%)**
	0.05	80.8%(2.6%)	81.4%(2.7%)	81.0%(3.1%)	**85.1%(2.7%)**
	0.0	82.6%(3.1%)	82.2%(2.1%)	83.2%(4.2%)	**87.5%(4.1%)**
LC	0.3	88.5%(2.1%)	89.0%(2.0%)	87.5%(2.8%)	**89.0%(1.1%)**
	0.1	**89.6%(0.9%)**	89.9%(2.3%)	89.2%(2.0%)	88.4%(1.4%)
	0.05	88.8%(0.9%)	**89.8%(1.7%)**	89.2%(1.7%)	88.6%(0.7%)
	0.0	91.7%(1.7%)	91.4%(1.4%)	91.7%(2.3%)	**94.1%(1.2%)**
C-T2D	0.3	76.7%(1.9%)	76.5%(1.5%)	74.7%(1.4%)	**76.9%(1.1%)**
	0.1	76.2%(2.3%)	75.3%(2.1%)	74.9%(1.9%)	**77.2%(1.7%)**
	0.05	76.8%(1.6%)	76.0%(2.8%)	77.5%(1.9%)	**77.5%(1.2%)**
	0.0	77.5%(1.9%)	76.6%(1.6%)	76.6%(1.6%)	**77.6%(1.6%)**
IBD	0.3	84.0%(5.0%)	85.8%(3.3%)	82.8%(5.8%)	**87.5%(2.5%)**
	0.1	84.9%(5.2%)	88.2%(3.6%)	87.2%(4.4%)	**91.2%(1.5%)**
	0.05	88.4%(3.5%)	**89.6%(2.4%)**	88.9%(3.2%)	88.4%(2.6%)
	0.0	92.0%(2.4%)	94.6%(1.7%)	95.5%(2.1%)	**97.4%(1.8%)**
Obesity	0.3	60.0%(3.5%)	60.8%(2.9%)	58.7%(1.4%)	**64.7%(2.9%)**
	0.1	59.1%(4.6%)	59.7%(3.1%)	61.7%(2.6%)	**64.7%(4.0%)**
	0.05	60.2%(2.7%)	59.2%(3.2%)	60.9%(4.1%)	**64.6%(4.3%)**
	0.0	61.5%(4.2%)	61.0%(3.3%)	62.2%(3.3%)	**67.7%(2.9%)**

simulated potential data perturbations that may occur in real-world scenarios. This setup allowed us to assess the stability and reliability of MSFT-Transformer when operating under noisy conditions.

From the overall experimental results, it is clear that the performance of all models declined to varying degrees as the noise level increased (from 0.0 to 0.3), indicating that noise had a significant impact on predictive capability. Notably, although increased noise negatively affected all models, the MSFT-Transformer demonstrated superior robustness, with smaller performance degradation and more stable predictions under high noise interference. Specifically, while the overall performance of all models decreased as noise intensified, our model consistently maintained optimal or competitive results among multimodal models. Even under a noise level of 0.3, it continued to deliver relatively accurate predictions compared with other methods, showcasing strong anti-noise capability. In addition, the *IBD* dataset exhibited more pronounced performance drops across all models upon noise injection. This is particularly evident in the *IBD* dataset, where severe class imbalance posed greater challenges for model learning.

Overall, despite the adverse effects of increased noise on model performance, MSFT-Transformer demonstrated remarkable stability and adaptability. Its consistent superiority under higher noise conditions further highlights its enhanced robustness compared with other approaches.

Ablation study

We also conduct an ablation study to examine the contributions of the fusion in three stages for MSFT-Transformer. We first define a multistage fusion baseline model with none of these novel fusion strategies and then add the three proposed modules at different stages, given their possible combinations. More specifically, we evaluate the performance of the following models under the same framework as shown in Table 5:

- *Baseline* denotes a conventional multistage fusion transformer model as the baseline, which initializes Z^0 randomly (from a normal distribution), does not align E_i^l with E_i^0, and employs conventional late fusion without Z^L.
- *Baseline + FAFEM* denotes the baseline model with FAFEM only, where Z^0 is learned from the features.
- *Baseline + AEFM* indicates the baseline model with AEFM only while Z^0 is random.
- *Baseline + IFDL* describes the baseline model with IFDL only, where $\hat{y}$ integrates Z^L while Z^0 is random.
- *Baseline + AEFM + IFDL* is the baseline model plus AEFM and IFDL, with Z^0 randomly initialized but E_i^l aligned with E_i^0 and $\hat{y}$ strengthened with Z^L.
- *MSFT-Transformer* is the final method when the baseline model is equipped with FAFEM, AEFM, and IFDL.

Several observations can be made from Table 5. (1) *Baseline + FAFEM versus Baseline*. There is a clear performance boost by enabling the learning of Z^0 from the input features with "Baseline + FAFEM" across all datasets. (2) *Baseline + AEFM versus Baseline*. "Baseline + AEFM" seems to reduce performance on two relatively smaller datasets, EW-T2D and Obesity, while demonstrating improved performance on LC and IBD, and maintaining consistent performance on C-T2D. It might indicate that integrating intra-modal alignment alone is insufficient to boost the overall performance of the model. (3) *Baseline + IFDL versus Baseline*. Similarly, "Baseline + IFDL" does not outperform the "Baseline" on most datasets for the same reason. (4) *Baseline + AEFM + IFDL versus Baseline*. "Baseline + AEFM + IFDL" outperforms "Baseline" on three datasets, EW-T2D, LC, and Obesity, and achieves improved performance compared with when AEFM and IFDL are used independently. (5) *MSFT-Transformer versus All other models*. "MSFT-Transformer," which incorporates all three key modules, FAFEM, AEFM, and IFDL, significantly outperforms all other models. While FAFEM alone provides the most significant performance boost, the combination of all three modules consistently outperforms other setups across the board. To conclude, FAFEM plays a crucial role, even with its straightforward MLP initialization of Z^0. It indicates that learned bottleneck features enhance cross-modal learning. AEFM and IFDL also make meaningful contributions, improving cross-modal alignment via E_i^l. The final bottleneck

Table 5. Ablation study on multistage fusion w.r.t. AUC (with the standard mean error in brackets)

Method	EW-T2D	LC	C-T2D	IBD	Obesity
Baseline	80.6%(4.6%)	92.7%(1.5%)	69.9%(2.6%)	90.8%(3.7%)	59.1%(2.9%)
Baseline + FAFEM	84.2%(2.4%)	93.3%(0.5%)	75.6%(1.6%)	95.7%(0.8%)	62.9%(3.1%)
Baseline + AEFM	79.1%(3.4%)	94.0%(1.3%)	69.9%(3.5%)	90.5%(3.7%)	61.8%(2.9%)
Baseline + IFDL	80.0%(5.4%)	92.0%(1.4%)	68.1%(2.9%)	89.6%(3.2%)	59.5%(3.8%)
Baseline + AEFM + IFDL	81.6%(2.7%)	94.0%(1.1%)	68.8%(4.2%)	89.9%(3.3%)	59.3%(2.4%)
MSFT-Transformer	**87.5%(4.1%)**	**94.1%(1.2%)**	**77.6%(1.6%)**	**97.4%(1.8%)**	**67.7%(2.9%)**

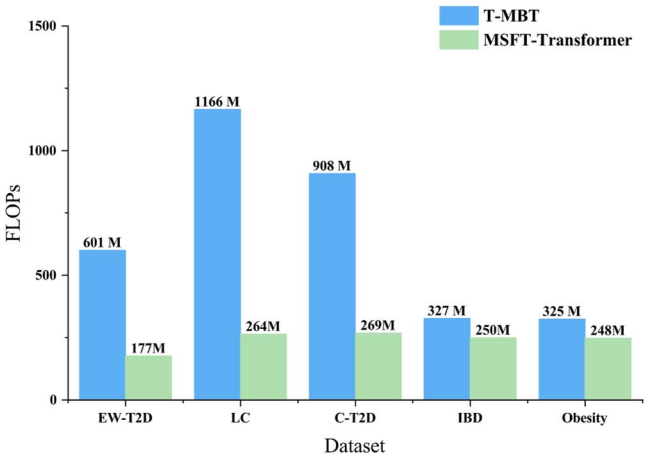

Figure 4. Computational complexity: T-MBT versus MSFT-Transformer across five datasets.

tokens do contribute to decision-making. Ultimately, the collaboration of all three key modules is essential to attaining stable and significant performance gains.

Computational complexity

Computational complexity is a fundamental problem in deep learning [51]. It not only impacts the inference speed of a model but also determines its feasibility and scalability under specific hardware resource constraints [52]. Given that T-MBT is the reference multistage fusion transformer model compared with our method, we provide a comparison of computational complexity between them. The computational complexity is quantified in terms of FLOPs (Floating Point Operations) using the *thop* Python library for both T-MBT and MSFT-Transformer, as shown in Fig. 4. Note that hyperparameter tuning is applied to obtain the optimal performance (as shown in Table 3). It shows that MSFT-Transformer is substantially more computationally efficient on the EW-T2D, LC, and C-T2D datasets.

Meanwhile, we observe that the number of FLOPs for the IBD and Obesity datasets does not show a significant disparity. For the *IBD* dataset, both models received relatively strong performance likely due to its pronounced inter-class feature separability. As a result, the computational efficiency differences between MSFT-Transformer and T-MBT is minimal. On the contract, for the *Obesity* dataset, the presence of relatively noise and limited data quality constrains both models' ability to extract discriminative features. This results in comparable representational complexity requirements between both models, particularly when approaching convergence (i.e. when further loss reduction becomes infeasible) during the training.

Feature importance analysis

In Table 6, we list these top five biomarkers (including bacteria and KO) according to their feature importance in each disease. We utilize the SHAP [53] to calculate the feature importance over five trails and take the average. To verify the importance of retained biomarkers, we implement case studies on five datasets by checking whether these biomarkers have been verified. If a biomarker has been reported to be associated with the disease, the source of the corresponding evidence will be listed in the table. For example, *Parabacteroides distasonis* can help alleviate T2D by repairing the gut barrier and reducing inflammation [54]. K00012 is associated with T2D through its role in glucose metabolism and kidney-related pathways [55]. The importance of other unlisted features can be found in our GitHub repository.

Conclusion

In this study, we presented the proposed MSFT-Transformer, a novel multimodal tabular transformer that applies improved fusion strategies across early-, mid-, and late-stage for disease prediction using metagenomic data. The proposed multistage fusion strategies include three modules in an MSFT-Transformer model: FAFEM, AEFM, and IFDL. These designs enforce (1) learning the initialized bottleneck tokens from the input features, (2) extra intra-modality alignment, which also leads to improved cross-modality learning, and (3) integrating bottleneck information into the decision-making mechanism. Following the detailed description of the design, we conducted in-depth experiments on five standard datasets comparing a few unimodal and multimodal approaches and several related works against MSFT-Transformer. Our results and analysis highlight a stable performance gain of the proposed model over the other unimodal & multimodal baseline approaches. Although not always outperforming methods engaging external knowledge, MSFT-Transformer reduces the cost of external data curation. Subsequent experiments further demonstrate the model's robustness to noisy and heterogeneous data compared with other multimodal baselines. Meanwhile, our ablation study illustrates the significant contributions of all three modules compared with a conventional multistage fusion transform, T-MBT, especially with reduced computational complexity. Furthermore, SHAP analysis revealed that the top taxonomic and functional features identified by MSFT-Transformer are largely consistent with previously reported disease-associated markers, demonstrating the model's effectiveness. It is also worth noting that, while the architecture proposed in this work is designed for taxonomic and functional data in tabular format, it can be easily extended to other omics data or data formats by simply replacing the feature extraction module to generate appropriate embeddings for those modalities. In summary, MSFT-Transformer can achieve both

Table 6. The SHAP is utilized to calculate the feature importance in our MSFT-Transformer

Disease	Biomarker	Importance	Evidence
EW-T2D	Dialister succinatiphilus	7.18	https://doi.org/10.3389/fendo.2021.814770
	Ruminococcus albus	4.19	https://doi.org/10.1093/femsre/fuad014
	Clostridium leptum	1.79	https://doi.org/10.1080/09637486.2021.1908964
	K07271 (lipopolysaccharide cholinephosphotransferase)	1.78	https://doi.org/10.21203/rs.3.rs-22813/v1
	Parabacteroides distasonis	1.31	https://doi.org/10.1186/s12915-023-01578-2
LC	Bacteroides dorei	1.67	https://doi.org/10.1128/spectrum.05349-22
	Bacteroides eggerthii	1.66	https://doi.org/10.1186/1471-230X-13-175
	Bifidobacterium catenulatum	1.59	https://doi.org/10.1007/s00248-011-9925-5
	Bacteroides coprocola	1.56	https://doi.org/10.3389/fmolb.2021.811399
	Clostridium symbiosum	1.52	https://doi.org/10.3350/cmh.2024.0349
C-T2D	K02911 (large subunit ribosomal protein L32)	1.37	–
	K01338 (ATP-dependent Lon protease [EC:3.4.21.53])	1.31	https://doi.org/10.1101/564492
	K20678 (dTDP-fucopyranose mutase [EC:5.4.99.59])	1.24	–
	Bacteroides vulgatus	1.21	https://doi.org/10.3389/fimmu.2017.01107
	K00012 (UDPglucose 6-dehydrogenase [EC:1.1.1.22])	1.21	https://doi.org/10.1016/j.ygeno.2022.110407
IBD	Bifidobacterium bifidum	0.36	https://doi.org/10.1016/j.clim.2006.11.005
	K00563 (23S rRNA (guanine745-N1)-methyltransferase [EC:2.1.1.187])	0.31	https://doi.org/10.3389/fmicb.2019.01902
	K14336 (23S rRNA (guanine748-N1)-methyltransferase [EC:2.1.1.188])	0.24	https://doi.org/10.1016/S1471-4914(03)00071-6
	K01193 (beta-fructofuranosidase [EC:3.2.1.26])	0.22	https://doi.org/10.7717/peerj.3698
	K06990 (MEMO1 family protein)	0.19	–
Obesity	Ruminococcus lactaris	2.79	https://doi.org/10.1038/nutd.2015.3
	Bilophila wadsworthia	2.69	https://doi.org/10.1038/s41467-018-05249-7
	Bilophila unclassified-	2.43	https://doi.org/10.1002/jsp2.70042
	Lachnospiraceae bacterium 8_1_57FAA	1.73	https://doi.org/10.2147/DDDT.S288011
	Ruminococcus bromii	1.31	https://doi.org/10.3389/fendo.2019.00941

The top five features over five trails are listed. The feature importance values represent the average across five independent experiments.

improved prediction and reduce the data and computational costs, implying its future potential in extended applications of disease prediction with metagenomic data. In future work, we plan to apply our approach to additional omics data and further evaluate the robustness of MSFT-Transformer under reduced data scenarios.

Key Points

- We introduced the MSFT-Transformer, a novel model that integrates taxonomic and functional information through a multistage fusion process, enhancing disease prediction without relying on domain-specific knowledge.
- The MSFT-Transformer employs a three-stage fusion process: early-stage fusion with a fusion-aware initialization strategy, mid-stage fusion with intra-modality alignment and cross-modality fusion, and late-stage fusion that combines modality-specific information for final decision-making.
- MSFT-Transformer demonstrates superior performance through extensive experiments on five metagenomic datasets, outperforming SOTA models across multiple metrics.

Conflict of interest: None declared.

Funding

We acknowledge support from the Fundamental Research Funds for the Central Universities (Grant No. JUSRP123035).

Data and code availability

The datasets used in this study are publicly available. Both our code and the accession links of the datasets can be found at https://github.com/WMGray/MTMF-Transformer.

References

1. Yamashiro Y. Gut microbiota in health and disease. *Ann Nutr Metab* 2018;**71**:242–6. https://doi.org/10.1159/000481627
2. Quince C, Walker AW, Simpson JT. *et al*. Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol* 2017;**35**:833–44. https://doi.org/10.1038/nbt.3935
3. Hugenholtz P, Tyson GW. Metagenomics. *Nature* 2008;**455**:481–3. https://doi.org/10.1038/455481a
4. Shen Y, Zhu J, Deng Z. *et al*. EnsDeepDP: an ensemble deep learning approach for disease prediction through metagenomics. *IEEE/ACM Trans Comput Biol Bioinform* 2022;**20**:986–98.
5. Mreyoud Y, Song M, Lim J. *et al*. MegaD: deep learning for rapid and accurate disease status prediction of metagenomic samples. *Life* 2022;**12**:669. https://doi.org/10.3390/life12050669
6. Jin S, Zeng X, Xia F. *et al*. Application of deep learning methods in biological networks. *Brief Bioinform* 2021;**22**:1902–17. https://doi.org/10.1093/bib/bbaa043
7. Min O, Zhang L. DeepMicro: deep representation learning for disease prediction based on microbiome data. *Sci Rep* 2020;**10**:6026.
8. Karlsson FH, Tremaroli V, Nookaew I. *et al*. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* 2013;**498**:99–103. https://doi.org/10.1038/nature12198
9. Qin N, Yang F, Li A. *et al*. Alterations of the human gut microbiome in liver cirrhosis. *Nature* 2014;**513**:59–64. https://doi.org/10.1038/nature13568

10. Qin J, Li Y, Cai Z. *et al*. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 2012;**490**:55–60. https://doi.org/10.1038/nature11450
11. Qin J, Li R, Raes J. *et al*. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 2010;**464**:59–65. https://doi.org/10.1038/nature08821
12. Le Chatelier, Nielsen T, Qin J. *et al*. Richness of human gut microbiome correlates with metabolic markers. *Nature* 2013;**500**:541–6. https://doi.org/10.1038/nature12506
13. Dhungel E, Mreyoud Y, Gwak H-J. *et al*. MegaR: an interactive r package for rapid sample classification and phenotype prediction using metagenome profiles and machine learning. *BMC Bioinformatics* 2021;**22**:1–12.
14. Reiman D, Metwally AA, Sun J. *et al*. PopPhy-CNN: a phylogenetic tree embedded architecture for convolutional neural networks to predict host phenotype from metagenomic data. *IEEE J Biomed Health Inform* 2020;**24**:2993–3001. https://doi.org/10.1109/JBHI.2020.2993761
15. Reiman D, Metwally AA, Sun J. *et al*. Meta-Signer: metagenomic signature identifier based on rank aggregation of features. *F1000Research* 2021;**10**:194. https://doi.org/10.12688/f1000research.27384.1
16. Sharma D, Paterson AD, Wei X. TaxoNN: ensemble of neural networks on stratified microbiome data for disease prediction. *Bioinformatics* 2020;**36**:4544–50. https://doi.org/10.1093/bioinformatics/btaa542
17. Chen X, Zhu Z, Zhang W. *et al*. Human disease prediction from microbiome data by multiple feature fusion and deep learning. *Iscience* 2022;**25**:104081. https://doi.org/10.1016/j.isci.2022.104081
18. Monshizadeh M, Ye Y. Incorporating metabolic activity, taxonomy and community structure to improve microbiome-based predictive models for host phenotype prediction. *Gut Microbes* 2024;**16**:2302076. https://doi.org/10.1080/19490976.2024.2302076
19. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:27–30. https://doi.org/10.1093/nar/28.1.27
20. Gene Ontology Consortium. The gene ontology (go) database and informatics resource. *Nucleic Acids Res* 2004;**32**:258D–261. https://doi.org/10.1093/nar/gkh036
21. Huerta-Cepas J, Szklarczyk D, Heller D. *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 2019;**47**:D309–14. https://doi.org/10.1093/nar/gky1085
22. Zhang Y-H, Guo W, Zeng T. *et al*. Identification of microbiota biomarkers with orthologous gene annotation for type 2 diabetes. *Front Microbiol* 2021;**12**:711244. https://doi.org/10.3389/fmicb.2021.711244
23. Liu Y, Zhang Y, Imoto S. Discovering microbe functionality in human disease with a gene-ontology-aware model. In: *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1873–80. Houston, TX, USA: IEEE, 2021.
24. Vaswani A, Shazeer N, Parmar N. *et al*. Attention is all you need. In: *Advances in neural information processing systems* (NIPS), Long Beach, CA, USA, 2017;**30**:5998–6008.
25. Peng X, Zhu X, Clifton DA. Multimodal learning with transformers: a survey. *IEEE Trans Pattern Anal Mach Intell* 2023;**45**:12113–32.
26. Nagrani A, Yang S, Arnab A. *et al*. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems* 2021;**34**:14200–13.
27. Acosta JN, Falcone GJ, Rajpurkar P. *et al*. Multimodal biomedical AI. *Nat Med* 2022;**28**:1773–84. https://doi.org/10.1038/s41591-022-01981-2
28. Lee SJ, Rho M. Multimodal deep learning applied to classify healthy and disease states of human microbiome. *Sci Rep* 2022;**12**:824. https://doi.org/10.1038/s41598-022-04773-3
29. Grazioli F, Siarheyeu R, Alqassem I. *et al*. Microbiome-based disease prediction with multimodal variational information bottlenecks. *PLoS Comput Biol* 2022;**18**:e1010050. https://doi.org/10.1371/journal.pcbi.1010050
30. Truong DT, Franzosa EA, Tickle TL. *et al*. Metaphlan2 for enhanced metagenomic taxonomic profiling. *Nat Methods* 2015;**12**:902–3. https://doi.org/10.1038/nmeth.3589
31. Pasolli E, Truong DT, Malik F. *et al*. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput Biol* 2016;**12**:e1004977. https://doi.org/10.1371/journal.pcbi.1004977
32. Chen S, Zhou Y, Chen Y. *et al*. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;**34**:i884–90. https://doi.org/10.1093/bioinformatics/bty560
33. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nat Methods* 2012;**9**:357–9. https://doi.org/10.1038/nmeth.1923
34. Xue C-X, Lin H, Zhu X-Y. *et al*. DiTing: a pipeline to infer and compare biogeochemical pathways from metagenomic and metatranscriptomic data. *Front Microbiol* 2021;**12**:698286. https://doi.org/10.3389/fmicb.2021.698286
35. Da Silva, Teterina A, Comelli EM. *et al*. Nonalcoholic fatty liver disease is associated with dysbiosis independent of body mass index and insulin resistance. *Sci Rep* 2018;**8**:1466. https://doi.org/10.1038/s41598-018-19753-9
36. Gorishniy Y, Rubachev I, Khrulkov V. *et al*. Revisiting deep learning models for tabular data. In: *Advances in Neural Information Processing Systems (NeurIPS), Virtual Event*, 2021;**34**:18932–43.
37. Devlin J, Lee M-WCK, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of NAACL-HLT, Minneapolis, MN, USA*, 2019;**1**:4171–86.
38. ChengguoYuan Y, Jin ZW, Wei F. *et al*. Learning bottleneck transformer for event image-voxel feature fusion based classification. In: *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 3–15. Xiamen, China: Springer, 2023. https://doi.org/10.1007/978-981-99-8429-91
39. Wang Z, Lequan Y, Ding X. *et al*. Shared-specific feature learning with bottleneck fusion transformer for multi-modal whole slide image analysis. *IEEE Trans Med Imaging* 2023;**42**:3374–83. https://doi.org/10.1109/TMI.2023.3287256
40. Budinger E, Heil P, Hess A. *et al*. Multisensory processing via early cortical stages: connections of the primary auditory cortical field with other sensory systems. *Neuroscience* 2006;**143**:1065–83. https://doi.org/10.1016/j.neuroscience.2006.08.035
41. Barnum G, Talukder SJ, Yue Y. On the benefits of early fusion in multimodal representation learning. In: *NeurIPS Workshop SVRHM*. Virtual Event, 2020.
42. Stahlschmidt SR, Ulfenborg B, Synnergren J. Multimodal deep learning for biomedical data fusion: a review. *Brief Bioinform* 2022;**23**:bbab569.
43. Li K, Peng J-X. Neural input selection—a fast model-based approach. *Neurocomputing* 2007;**70**:762–9. https://doi.org/10.1016/j.neucom.2006.10.011
44. Xu H, Zeng R, Wu Q. *et al*. Cross-modal relation-aware networks for audio-visual event localization. In: *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 3893–901. WA, USA: Virtual Event / Seattle, 2020.

45. Zhu J, Xia Y, Wu L. *et al*. Incorporating bert into neural machine translation. In: *International Conference on Learning Representations (ICLR)*. Addis Ababa, Ethiopia, 2019.
46. Chen C-FR, Fan Q, Panda R. CrossViT: cross-attention multi-scale vision transformer for image classification. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 347–56. Montreal, QC, Canada, 2021.
47. Haykin S. *Neural Networks: A Comprehensive Foundation*. Upper Saddle River, NJ, United States: Prentice Hall PTR, 1994.
48. Ho TK. Random decision forests. In: *Proceedings of 3rd International Conference on Document Analysis and Recognition (ICDAR)*, Vol. **1**, pp. 278–82. Montreal, Canada: IEEE, 1995.
49. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;**20**:273–97.
50. Pedregosa F, Varoquaux G, Gramfort A. *et al*. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;**12**:2825–30.
51. Xia H, Chu L, Pei J. *et al*. Model complexity of deep learning: a survey. *Knowl Inf Syst* 2021;**63**:2585–619.
52. Canziani A, Paszke A, Culurciello E. An analysis of deep neural network models for practical applications. arXiv preprint arXiv:1605.07678. 2016;1–7.
53. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: *Advances in neural information processing systems (NeurIPS)*, pp. 4765–74. Long Beach, CA, USA, 2017.
54. Liu D, Zhang S, Li S. *et al*. Indoleacrylic acid produced by parabacteroides distasonis alleviates type 2 diabetes via activation of AhR to repair intestinal barrier. *BMC Biol* 2023;**21**:90. https://doi.org/10.1186/s12915-023-01578-2
55. Park S, Kim O-H, Lee K. *et al*. Plasma and urinary extracellular vesicle micrornas and their related pathways in diabetic kidney disease. *Genomics* 2022;**114**:110407. https://doi.org/10.1016/j.ygeno.2022.110407

Briefings in Bioinformatics, 2025, **26(3)**, bbaf217
https://doi.org/10.1093/bib/bbaf217
Problem Solving Protocol