Information Fusion 117 (2025) 102890

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

OmniFuse: A general modality fusion framework for multi-modality learning on low-quality medical data

Yixuan Wu [a], Jintai Chen [b,*], Lianting Hu [c], Hongxia Xu [a], Huiying Liang [d,e,**], Jian Wu [a,***]

[a] State Key Laboratory of Transvascular Implantation Devices of The Second Affiliated Hospital School of Medicine and School of Public Health and Liangzhu Laboratory, Zhejiang University, Hangzhou, China
[b] AI Thrust, Information Hub, HKUST (Guangzhou), Guangzhou, China
[c] The Data Center, Wuhan Children's Hospital (Wuhan Maternal and Child Healthcare Hospital), Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China
[d] Medical Big Data Center, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, China
[e] Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Guangzhou, China

ARTICLE INFO

MSC:
62P10

Keywords:
Multi-modal fusion
Low-quality medical data
Data imputation

ABSTRACT

Mirroring the practice of human medical experts, the integration of diverse medical examination modalities enhances the performance of predictive models in clinical settings. However, traditional multi-modal learning systems face significant challenges when dealing with low-quality medical data, which is common due to factors such as inconsistent data collection across multiple sites and varying sensor resolutions, as well as information loss due to poor data management. To address these issues, in this paper, we identify and explore three core technical challenges surrounding multi-modal learning on low-quality medical data: (i) the absence of informative modalities, (ii) imbalanced clinically useful information across modalities, and (iii) the entanglement of valuable information with noise in the data. To fully harness the potential of multi-modal low-quality data for automated high-precision disease diagnosis, we propose a general medical multi-modality learning framework that addresses these three core challenges on varying medical scenarios involving multiple modalities. To compensate for the absence of informative modalities, we utilize existing modalities to selectively integrate valuable information and then perform imputation, which is effective even in extreme absence scenarios. For the issue of modality information imbalance, we explicitly quantify the relationships between different modalities for individual samples, ensuring that the effective information from advantageous modalities is fully utilized. Moreover, to mitigate the conflation of information with noise, our framework traceably identifies and activates lazy modality combinations to eliminate noise and enhance data quality. Extensive experiments demonstrate the superiority and broad applicability of our framework. In predicting in-hospital mortality using joint EHR, Chest X-ray, and Report dara, our framework surpasses existing methods, improving the AUROC from 0.811 to 0.872. When applied to lung cancer pathological subtyping using PET, CT, and Report data, our approach achieves an impressive AUROC of 0.894.

1. Introduction

Our perception of the world relies on multiple sensory modalities such as touch, sight, hearing, smell, and taste. Even when some sensory signals are unreliable, humans are adept at extracting useful information from imperfect multimodal inputs, thereby constructing a coherent understanding of events. With advancements in sensory technology [1,2], we can now easily gather diverse forms of data for analysis. To fully harness the value of each modality, multimodal fusion has emerged as a promising approach to achieve precise and reliable predictions by integrating all available cues for downstream analytical tasks [3–8].

* Corresponding author.
** Corresponding author at: Medical Big Data Center, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, China.
*** Correspondence to: State Key Laboratory of Transvascular Implantation Devices of The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China.
E-mail addresses: wyx_chloe@zju.edu.com (Y. Wu), jtchen721@gmail.com (J. Chen), lianghuiying@hotmail.com (H. Liang), wujian2000@zju.edu.cn (J. Wu).

https://doi.org/10.1016/j.inffus.2024.102890
Received 16 August 2024; Received in revised form 23 November 2024; Accepted 16 December 2024
Available online 24 December 2024

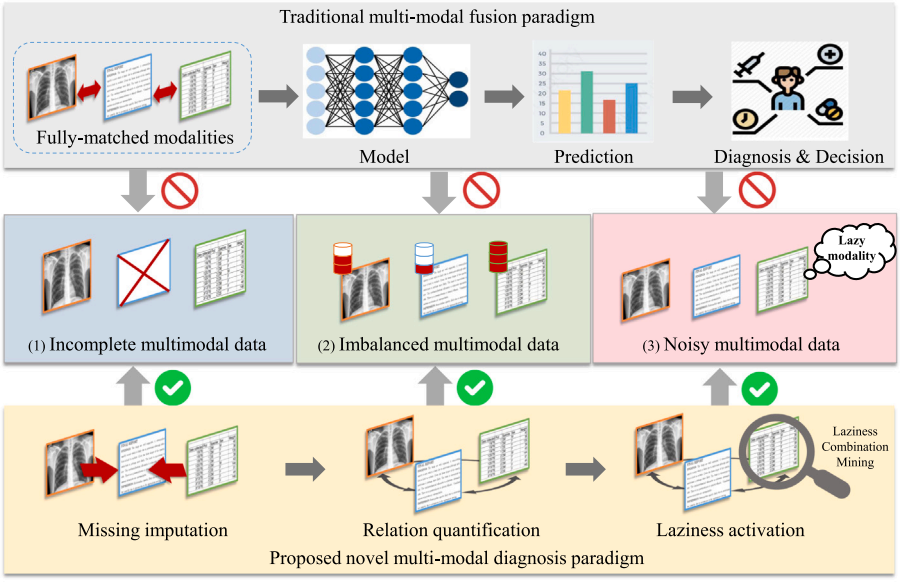

Fig. 1. Three major challenges in low-quality multimodal medical prediction scenarios: Incomplete multimodal data, imbalanced multimodal data, as well as noisy multimodal data. Our novel framework can simultaneously address these three challenges by performing missing imputation for any missing combination, quantifying relationships between modalities to determine their predictive contributions, and identifying and activating "lazy modalities".

However, there is growing recognition that widely-used AI models are often misled by spurious correlations and biases present in low-quality data. In real-world clinical settings, the quality of different modalities frequently varies due to factors such as inconsistent data collection across sites, varying sensor resolutions, and data degradation or loss from poor data management. Recent empirical and theoretical studies have shown that conventional multimodal fusion techniques often struggle with low-quality data, especially when faced with issues like missing, imbalanced, or noisy data [9–11]. Crucially, there is no existing framework that addresses these challenges — missing, imbalanced, and noisy data — within a single, unified approach. Current methods are typically designed to handle one specific issue, lacking a comprehensive strategy that can adaptively manage all three (see Fig. 1). This limitation hinders their practicality in the complex and variable conditions encountered in real-world medical applications.

To address these limitations and advance robust and generalized multimodal learning in clinical applications, we identify three characteristics of low-quality multimodal data and focus on the unique challenges inherent in multimodal machine fusion for disease diagnosis and prediction:

(1) **Incomplete multimodal data.** Medical data often experience intermittent absence or incompleteness of some modalities [9]. Due to the variability in patient conditions, not all diagnostic tests are performed for every patient, leading to extreme cases of modality missingness. Moreover, patients even with the same disease may choose different medical examinations producing incomplete multimodal data. Traditionally, multimodal learning systems have struggled with the absence of one or more modalities, often relying on static data imputation then fusion methods that do not account for the dynamic nature of data streams of specific patient sample in real clinical applications.

(2) **Imbalanced multimodal data.** Compared to other domains like remote sensing [12], where different modalities (e.g., various types of images) have relatively small modality gaps, medical multimodal data encompass a broader range of modality gaps, including images (e.g., CT, MRI, X-ray), structured data (e.g., EHR), and unstructured textual data (e.g., clinical reports). This kind of imbalance results in the problem of imbalanced clinically useful information across modalities, making it challenging to quantify the contribution of each modality during fusion to the predictive outcome. This often leads to a counterintuitive fact that the performance of multi-modal fusion predictive models can be inferior to that of dominant uni-modal models. Our preliminary experiments in Fig. 2 also reveal that the predictive accuracy of combined modalities is lower than that of single modality predictions. We define this phenomenon as "modality laziness" issue, where simply combining multi-modal data without additional intervention leads to a significant degradation in performance.

(3) **Noisy multimodal data.** Another underlying reason contributing to the "modality laziness" issue is the intermixing of noise within certain uninformative modalities. High-dimensional multimodal data tend to contain complex noise. In the process of collecting medical data, it is easy to include noise, such as metal artifacts produced during medical imaging. When noise becomes interwoven with the effective information from these modalities, it hinders the predictive power of the valuable information, ultimately degrading the overall prediction performance.

Toward addressing these increasingly important but sometimes overlooked issue of "modality laziness" in multimodal fusion, this paper proposes a traceable prediction framework for low-quality medical data in a dynamic trustworthy fusion manner. A key feature of our framework is its capability to handle any extreme combination of missing modalities. Moreover, our proposed framework introduces an innovative adaptive modal fusion technique, Dynamic Weighted Fuse (DWFuse), based on the explicit quantification of the relationships between different modalities. This approach ensures that the effective information from the advantageous modalities is fully utilized for specific samples. Additionally, we incorporate a novel training strategy, Traceble Laziness Activation (TLA), for traceably mining the lazy modality combinations and activate the lazy ones at a granular level. This allows for the explicit identification and activation of "lazy modalities", thereby unleashing the potential of multi-modal data sources for achieving high-precision disease prediction.

The experimental validation of our framework across diverse datasets, such as MIMIC-III [13] and MIMIC-IV [14] for CXR images, clinical reports, and structured EHR modalities, along with a proprietary lung cancer pathological subtyping prediction dataset for CT and PET images and clinical reports. These datasets encompass a variety of modalities (e.g., X-ray, CT, PET, textual reports, and EHR) and languages (e.g., English and Chinese clinical reports), demonstrating the robustness, versatility, and generalizability of our framework across different data types and linguistic contexts. The framework's capability

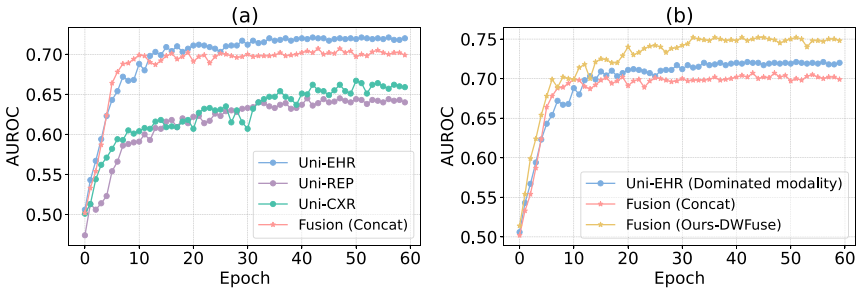

Fig. 2. Performance of the uni-modality, fused multi-modality in a concatenated manner, and fused multi-modality with our proposed DWFuse module on the validation set of the in-hospital mortality prediction task. (a) One can see that the dominated uni-modality is EHR modality, while fusing all modalities by concatenation leads to inferior performance. (b) When utilizing our proposed DWFuse module, the performance of multi-modality prediction is better than single dominated modality.

to dynamically adapt to evolving data scenarios and provide traceable, reliable predictive outputs positions it as a valuable tool in advancing multimodal learning.

2. Related work

2.1. Incomplete multimodal learning

In real-world clinical applications, collected multimodal data are often incomplete, with some patients missing certain modalities [9,15–17]. For example, in Alzheimer's Disease diagnosis [15], although combining data from multiple modalities, such as magnetic resonance imaging (MRI) scans, positron emission tomography (PET) scans, and cerebrospinal fluid (CSF) information, can enhance diagnostic accuracy, the high cost of PET scans and the invasive nature of CSF tests may lead some patients to decline these examinations. From the perspective of handling missing data, existing methods can be categorized into two groups: imputation-based [18–23] and imputation-free [24–27] incomplete multimodal learning. Imputation-based methods often employ techniques like zero imputation [20] and mean value imputation [21] in the early stages. Additionally, many learning-based imputation methods [22,23] have been developed to fill in missing modalities for specific tasks and samples. However, these methods often lead to biased results due to the lack of reliable information, and the imputed values may degrade model performance rather than improve it, particularly in complex medical scenarios. Moreover, imputation-based methods are generally unsuitable for extreme missing scenarios, where performance tends to deteriorate significantly. On the other hand, imputation-free methods focus solely on leveraging the information from available modalities without attempting to fill in the missing data. While this avoids the biases associated with imputation, these methods still struggle to fully leverage the partial information and often lack the ability to dynamically adapt to situations with missing data, limiting their overall effectiveness.

2.2. Imbalanced multimodal learning

Recent studies have highlighted the challenge of imbalanced multimodal learning [10,28–30], where multimodal models often prioritize specific modalities, limiting their overall performance. This imbalance arises because each modality has unique data sources and forms, leading to varying quality levels. Some modalities provide richer, more direct information, while others offer less. Given the greedy nature of deep neural networks [30], multimodal models tend to rely heavily on the high-quality modalities with sufficient target-related information, neglecting the less informative ones. It is worth noting that in practical clinical scenarios, the quality of each modality dynamically changes with respect to different patients and tasks [31]. To address this issue, various methods have been proposed [10,28,29], focusing on different aspects such as learning objectives, optimization processes, and data augmentation. Learning-objective-based methods [28] introduce additional loss functions to counteract the model's modality preference. Optimization-based methods [32] concentrate on the back-propagation stage, adjusting the magnitude and direction of unimodal gradients to enhance the learning of lower-quality modalities. Data-augmentation-based methods [30,33,34] aim to improve the lower-quality modalities at the data input stage by enhancing their information content. However, many existing methods require the introduction of additional neural modules, which significantly complicates the training process and increases computational demands. This added complexity often makes these methods impractical, especially in resource-constrained settings like healthcare. Additionally, data-augmentation-based methods are also not immune to these challenges, as their effectiveness is heavily dependent on the initial quality of the data. If the original data is highly degraded, augmentation efforts may fail to produce meaningful improvements. Furthermore, such methods risk introducing artificial biases, as augmentation can create unrealistic patterns or artifacts that ultimately degrade the model's ability to generalize effectively to real-world cases.

2.3. Noisy multimodal learning

In real-world scenarios, collecting high-quality multimodal data is inherently challenging due to the inevitable presence of noise [11,35–37]. This multimodal noise can be broadly categorized based on its source: (1) modality-specific noise [37–40], which results from issues like sensor errors, environmental factors, or transmission errors specific to each modality. For example, in medical imaging, electronic noise in sensors can lead to loss of detail, and metal artifacts are often present [41]. (2) Cross-modal noise [42–46], which arises from weakly aligned or unaligned multimodal pairs and can be considered semantic-level noise. In medical contexts, this misalignment occurs naturally because data from different modalities are often captured at different times while the patient's condition evolves [14,47]. Fortunately, leveraging the correlations between multiple modalities or optimizing the utilization of multimodal data can significantly aid in the effective fusion of noisy data. By exploiting inter-modal correlations, multimodal data can identify and mitigate potential noise, thereby enhancing the overall data quality and predictive accuracy [36,48]. However, a key shortcoming of methods leveraging modal correlations is their reliance on the assumption that data from all modalities are complete. In real-world scenarios, missing data is common, which prevents the model from accurately capturing cross-modal relationships. This limits the effectiveness of correlation-based fusion, leading to unreliable noise reduction and suboptimal predictive performance when modalities are incomplete.

2.4. Multi-modal fusion in medical scenarios

Several studies [16,17,49–55] have investigated the fusion of multimodal medical data for various applications. While these studies highlight the positive impact of using multiple modalities on downstream

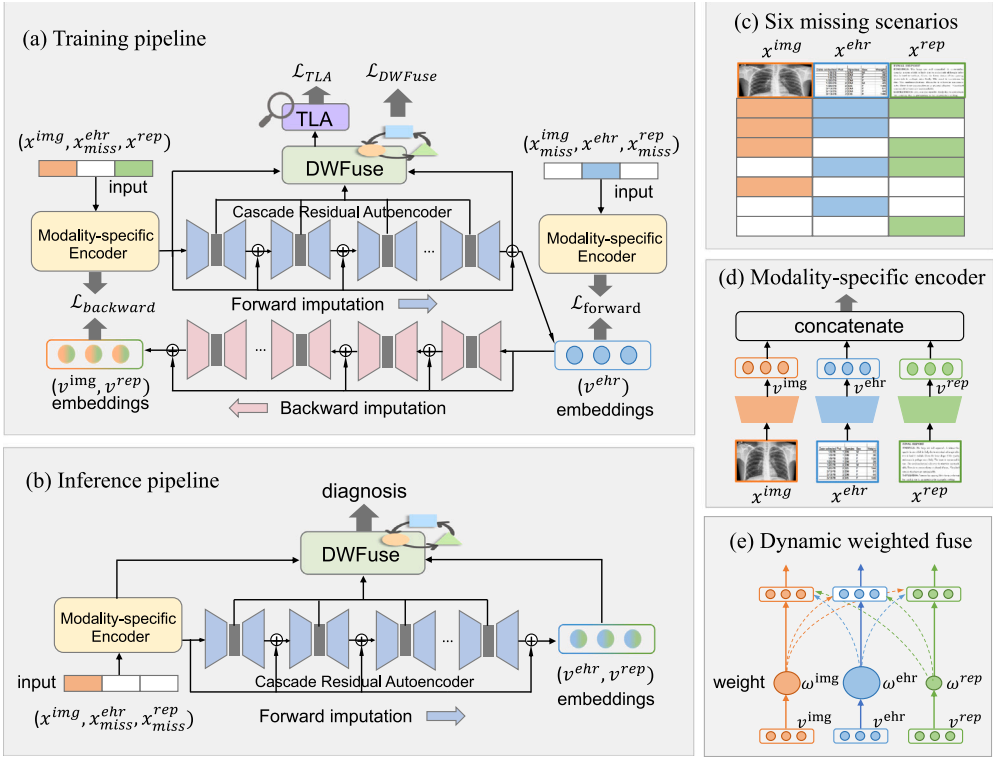

Fig. 3. Overview of the proposed network architecture: (a) Training stage: Inputs with arbitrary missing modality patterns are first encoded using modality-specific encoders, which are then fed into an imputation network to perform forward imputation of the missing modalities. The completed modalities are dynamically weighted and fused for joint prediction. Additionally, the embeddings of the imputed missing modalities undergo backward imputation to regenerate the embeddings of the initially available modalities, thereby facilitating dual-direction supervision through cycle consistency. (b) Inference stage: Only forward imputation is performed to fill in the missing modalities, followed by joint prediction. (c) Our method can handle any combination of missing modalities. (d) Modality-specific encoder: Each modality is independently encoded to obtain its corresponding embeddings. (e) Dynamic weighted fuse: This component quantifies the relationships among imbalanced modalities, and dynamically weights and aggregates them.

performance, many of them curate datasets for specific tasks and operate under the assumption that images and clinical features are paired. For instance, some works focus on tasks related to cancer recurrence prediction [54], lesion detection [53], patient survival prediction [55], as well as predicting the progression of Alzheimer's disease [16,17] and prognostication for COVID-19 patients [52]. Among them, MedFuse [50] and TriMF [51] are two studies closely related to our work. MedFuse is an LSTM-based multimodal fusion method that integrates clinical time-series data with chest X-ray images. However, it is limited to two modalities — EHR and images — whereas our framework incorporates an additional textual report modality, making it more generalizable. Moreover, our approach demonstrates superior performance in MIMIC in-hospital mortality prediction, achieving an AUC score of 0.859 compared to 0.845 for MedFuse (as shown in Table 3). TriMF, on the other hand, introduces an efficient multimodal fusion architecture designed to be robust to missing modalities. However, it uses an imputation-free fusion method, which often results in lower accuracy in severe missing scenarios, as it can lead to data and prediction biases. In contrast to these methods, which are primarily evaluated on public datasets, our framework undergoes additional validation on a self-collected clinical dataset with Chinese textual reports, demonstrating both its generalizability and applicability in real-world, multilingual medical settings. This evaluation underscores the robustness of our approach across diverse data environments.

3. Methodology

We start this section by providing an overview of the proposed framework (see Fig. 3) and formulation of the research problem. Then we detail the structure of each module and the entire training pipeline.

3.1. Problem formulation

Given a patient p, it has multiple information sources from different modality inputs x^m, $m \in \{1, \ldots, M\}$, where M is the number of modalities. For example, we use $x = (x^{img}, x^{ehr}, x^{rep})$ to represent the raw input of three modalities, where x^{img} represents the medical imaging modality, x^{ehr} represents the modality of structured electronic health record (ehr) data, and x^{rep} represents the clinical textual report belonging to the same patient. We denote the patient's disease as y, $y \in \{1, \ldots, C\}$, where C is the number of disease categories. Our proposed method aims to predict the disease category y for every patient p under any combination of missing modality scenario.

3.2. Modality-specific encoders

One of the primary sources of modality heterogeneity arises from the fact that different modalities actually reside in different input and embedding spaces, posing challenges in devising a unified encoder capable of handling all modalities uniformly. Additionally, another distinguishing factor lies in the target space, as we do not presume that each modality must be associated with the same set of labels. Given these challenges, modality-specific encoders offer significant advantages in terms of scalability and flexibility. By allowing the independent addition of new modalities without affecting the performance of existing ones, this modular design ensures that new encoders can be trained specifically for each new data type. Consequently, we begin by establishing modality-specific encoders e^m for each modality m individually to get the corresponding embeddings, as:

$$v^m = e^m(x^m), \tag{1}$$

where x^m is the raw input of modality m, v^m denotes the encoded embeddings of modality m.

Table 1
The six modality missing scenarios and their detailed available-missing pairs.

No.	(Available, Missing)	Different missing scenarios
1	$(x^{\text{img}}), (x^{\text{ehr}}, x^{\text{rep}})$	$(x^{\text{img}}, x^{\text{ehr}}_{\text{miss}}, x^{\text{rep}}_{\text{miss}}), (x^{\text{img}}_{\text{miss}}, x^{\text{ehr}}, x^{\text{rep}})$
2	$(x^{\text{ehr}}), (x^{\text{img}}, x^{\text{rep}})$	$(x^{\text{img}}_{\text{miss}}, x^{\text{ehr}}, x^{\text{rep}}_{\text{miss}}), (x^{\text{img}}, x^{\text{ehr}}_{\text{miss}}, x^{\text{rep}})$
3	$(x^{\text{rep}}), (x^{\text{img}}, x^{\text{ehr}})$	$(x^{\text{img}}_{\text{miss}}, x^{\text{ehr}}_{\text{miss}}, x^{\text{rep}}), (x^{\text{img}}, x^{\text{ehr}}, x^{\text{rep}}_{\text{miss}})$
4	$(x^{\text{img}}, x^{\text{ehr}}), (x^{\text{rep}})$	$(x^{\text{img}}, x^{\text{ehr}}, x^{\text{rep}}_{\text{miss}}), (x^{\text{img}}_{\text{miss}}, x^{\text{ehr}}_{\text{miss}}, x^{\text{rep}})$
5	$(x^{\text{img}}, x^{\text{rep}}), (x^{\text{ehr}})$	$(x^{\text{img}}, x^{\text{ehr}}_{\text{miss}}, x^{\text{rep}}), (x^{\text{img}}_{\text{miss}}, x^{\text{ehr}}, x^{\text{rep}}_{\text{miss}})$
6	$(x^{\text{ehr}}, x^{\text{rep}}), (x^{\text{img}})$	$(x^{\text{img}}_{\text{miss}}, x^{\text{ehr}}, x^{\text{rep}}), (x^{\text{img}}, x^{\text{ehr}}_{\text{miss}}, x^{\text{rep}}_{\text{miss}})$

3.3. Missing modality imputation

In medical scenarios, the issue of missing data is both prevalent and significant. Even for the same disease, different clinicians may order distinct diagnostic tests, leading to varied data sources. Nevertheless, we posit that there exist underlying connections between modalities, which collectively reflect a patient's health status and disease progression. Based on this hypothesis, we introduce an imputation module for missing modalities, designed to infer the embeddings of unavailable modalities from those of available ones. A notable advantage of our approach is its adaptability to any extreme modality missing situation, as summarized in Table 1.

As shown in Fig. 3, the input consists a concatenated triplet format of both available modalities and missing modalities. The multimodal embeddings of these cross-modality triplet can be represented as (taking the EHR modality missing condition as an example):

$$\upsilon = \text{concat}(\upsilon^{img}, \upsilon^{ehr}_{miss}, \upsilon^{rep}), \tag{2}$$

$$\hat{\upsilon} = \text{concat}(\upsilon^{img}_{miss}, \upsilon^{ehr}, \upsilon^{rep}_{miss}), \tag{3}$$

where υ^{img}_{miss}, υ^{ehr}_{miss}, and υ^{rep}_{miss} represent the modality-specific embedding when the corresponding modality is missing, which is produced by the corresponding modality encoder with input zero vectors.

Specifically, we utilize the Cascade Residual Autoencoder (CRA) structure, which surpasses the standard autoencoder in learning capacity and stability, the module integrates a sequence of Residual Autoencoders (RAs) for enhanced representation learning. To further enhance the model, we incorporate cycle consistency learning within a dual-network architecture to enable bidirectional forward (from available to missing modalities) and backward (from missing to available modalities). The primary reason for utilizing cycle consistency learning within our dual-network architecture is to ensure consistency and reliability when imputing missing modalities. Specifically, a CRA model comprising K RAs, denoted as φ_k for $k = 1, \ldots, K$. For $k > 1$, each RA's output, Δz_k, is iteratively refined by adding the cumulative outputs of preceding RAs to the input:

$$\begin{cases} \Delta z_k = \varphi_k(\upsilon), k = 1, \\ \Delta z_k = \varphi_k\left(\upsilon + \sum_{j=1}^{k-1} \Delta z_j\right), k > 1. \end{cases} \tag{4}$$

In this context, υ represents the cross-modality triplet embedding derived from available modalities. Taking the scenario where the EHR modality is absent while the IMG and REP modalities are available, as an example, the forward imputation predicts the missing EHR modality's embedding from IMG and REP modalities. The generated forward imputation embedding, υ', is the sum of υ and the outputs of all RAs. The backward imputation then predicts the embedding of the available modalities i.e., IMG and REP, υ'', from the forward imputed embedding, υ':

$$\upsilon' = \text{forward}(\upsilon) = \upsilon + \sum_{k=1}^{K} \Delta z_k, \tag{5}$$

$$\upsilon'' = \text{backward}(\upsilon') = \upsilon' + \sum_{k=1}^{K} \Delta z_k. \tag{6}$$

The training objective for missing modality imputation includes forward imputation loss $L_{forward}$ and backward imputation loss $L_{backward}$:

$$L_{forward} = \left\| \upsilon' - \upsilon \right\|_2^2, \tag{7}$$

$$L_{backward} = \left\| \upsilon - \upsilon'' \right\|_2^2, \tag{8}$$

where υ' and υ'' are imputed embeddings in the forward and backward imputation process, respectively. υ is the ground-truth embeddings extracted by corresponding modality-specific encoder.

3.4. Dynamic weighted fuse

In traditional multimodal prediction methods, all modalities are typically aggregated without considering their individual contributions. However, our observations indicate that not all modalities are equally influential in the prediction process for a specific patient. Specifically, within the low-quality data of different modalities for specific patients, there exist **key modalities** that are informative, **neutral modalities** that are less informative, and **lazy modalities** that even hinder the predictive performance. Our pre-experiments in Fig. 2 imply that the inclusion of lazy modalities without any intervention can result in significant performance degradation.

To mitigate the hindrance caused by lazy modalities, we introduce a mechanism, Dynamic Weighted Fuse (DWFuse), that reduces their influence by down-weighting their contribution when more informative modalities are available for a given patient. This ensures that the lazy modalities do not interfere with the predictive guidance of the key modalities. A modality is deemed reliable (or unreliable) based on whether it assigns a high (or low) probability to the correct category, with higher probabilities indicating more informative signals and greater confidence. Therefore, we define a down-weighting factor, ω_i, to dynamically adjust each modality's contribution by considering the confidence levels of other modalities, as follows:

$$p^m = g^m(\upsilon^m), \tag{9}$$

$$\omega^m = [\prod_{n \neq m} (1 - p^n)]^{\beta/(M-1)}, \tag{10}$$

where $g^m(\cdot)$ denotes the disease classifier for modality m that consists of several fully-connected layers, and p^m is the prediction of modality m. M is the number of modalities, and β is a hyperparameter determining the down-weighting intensity and are chosen by cross-validation. β controls the suppression strength: higher values intensify the effect, and vice versa. This scaling factor ω^m represents the average prediction quality of the remaining modalities n ($n \neq m$). This factor approaches 0 as some p_n approach 1, indicating when other modalities (excluding modality m) confidently predict the correct category, thereby minimizing the cost on the current modality (p_m). This approach ensures that modality weighting responds to the multimodal context rather than being fixed on a predefined accuracy metric, enhancing prediction robustness.

We also collect the latent vectors of each autoencoder in the forward imputation module and concatenate them together to form the joint multimodal embedding, which inherently captures interactions between modalities. This joint embedding is designed to facilitate cross-modal information interaction by integrating shared information while preserving individual modality-specific details. Based on the joint multimodal embedding, we calculate the joint probability distribution o as:

$$o = g_{ra}(\text{concat}(c_1, c_2, \ldots, c_K)), \tag{11}$$

where $g_{ra}(\cdot)$ denotes the disease classifier for joint multimodal embedding, and c_k is the latent vector of the autoencoder in the kth RA, $k = \{1, 2, \ldots, K\}$.

Consequently, the training objective for predicting the correct disease category is set as:

$$L_{enc} = -\sum_{m=1}^{M} H(\omega_i p_i, q), \tag{12}$$

$$L_{ra} = -H(o, q), \tag{13}$$

$$L_{DW\,Fuse} = L_{enc} + \alpha L_{ra}, \tag{14}$$

where $H(\cdot)$ is the CrossEntropy Loss, q is the true distribution of disease category label, and α is the weighting hyperparameter for L_{ra}.

3.5. Traceable laziness activation

In the previous section of DWFuse, we dynamically mitigate the side effects of less informative modalities by leveraging their strengths. However, we believe that the fundamental reason certain modalities exhibit "laziness" is due to their inherent noise, which may be introduced during data generation or collection processes, such as metal artifacts in medical imaging. This noise prevents the modality from fully realizing its predictive potential. Based on this, to unleash the potential of the "lazy modalities" and eliminate their noisy components, we propose a Traceable Laziness Activation (TLA) training strategy. This strategy first identifies the combinations of lazy modalities and then activates them. Assuming there are a total of M modalities, there are $2^M - 1$ possible modality combinations, denoted as $C = \{\delta_i \mid i \in 2^M - 1\}$. Specifically, TLA first proposes a contrastive ranking strategy to mine the lazy modality combinations δ_{lazy}. Compared to simply taking the single modality as the lazy one, our method takes a more comprehensive approach by considering combinations of multiple modalities, leading to more precise and context-aware mining of underperforming modality groups. This combination-based strategy enables TLA to better capture the interactions and dependencies between modalities, ensuring a more effective activation of lazy modality combinations. Then, TLA calculates the prediction loss for the lazy modality combinations, guiding the network to pay more attention to them. The process consists of two steps: (a) mining the lazy modality combination, and (b) calculating L_{TLA} to apply targeted regularization on the lazy modality, thereby achieving laziness activation.

(a) Laziness Combination Mining. This step is based on the assumption that DNNs tend to first memorize simple and informative examples before overfitting hard and noisy examples. Therefore, we first mine the strong modality via this memorization effect. Then we determine the remaining combinations of modalities as the lazy ones. Specifically, after each training epoch, compute the predicted output $\hat{y}_i$ for each modality combination δ_i, $i \in 2^M - 1$, and subsequently calculate their distances from the ground truth label y. Given that the neural networks tend to initially memorize samples featuring strong modalities, the modality combination δ_i exhibiting the greatest distance from y can be identified as the lazy modality combination. To enhance the robustness of lazy modality combinations against neural network learning randomness, TLA first computes sample-wise distances then integrates them into the combination-wise distances, as:

$$\delta_{\text{lazy}} = \arg\max_{\delta_i} \left(\sum_i D(\hat{y}_i, y) \right). \tag{15}$$

(b) Laziness Activation. First, we calculate the lazy combination mask $Mask \in \mathbb{R}^b$, as:

$$Mask(s) = \begin{cases} \text{FALSE}, & \text{if } \delta_i(s) \neq \delta_{\text{lazy}} \\ \text{TRUE}, & \text{if } \delta_i(s) = \delta_{\text{lazy}} \end{cases} \tag{16}$$

where b is the batch size, s is the index of patient p_s, $s \in [0, b-1]$, and $\delta_i(s)$ is the modality combination for the sth patient in this mini-batch. Then, the L_{TLA} is defined as follows:

$$L_{TLA} = H(\hat{y}[Mask],\ y[Mask]), \tag{17}$$

where $H(\cdot)$ is the CrossEntropy Loss, y is the ground truth label for predicted label $\hat{y}$. The $[\cdot]$ operator denotes the index operator.

3.6. Training objective

We combine all the losses into the joint objective function as below to jointly optimize the model parameters:

$$L = L_{forward} + \lambda_1 L_{backward} + \lambda_2 L_{DW\,Fuse} + \lambda_3 L_{TLA}, \tag{18}$$

where λ_1, λ_2 and λ_3 are weighting hyperparameters for each loss item.

4. Experimental setup

4.1. Datasets and tasks

We employ two diverse multi-modal datasets to address three distinct clinical tasks. The datasets used are:

- **MIMIC III and IV Datasets** [13,14]: Involves **in-hospital mortality prediction** and **phenotype classification** tasks. In-Hospital Mortality Prediction is a binary classification task aimed at predicting in-hospital mortality after the first 48 h of an ICU admission. Phenotype Classification is a multi-label classification task to predict the presence of 25 chronic, mixed, and acute care conditions.
- **Lung Cancer Pathological Subtyping Dataset**: Involves a multi-class classification task for **lung cancer pathological subtyping**. The pathological subtypes of lung cancer include six categories: Atypical Adenomatous Hyperplasia (AAH), Adenocarcinoma In Situ (AIS), Minimally Invasive Adenocarcinoma (MIA), Invasive Adenocarcinoma (IA) Grade 1, IA Grade 2, and IA Grade 3. Clinically, the risk of invasion is classified as low (L1: AAH, AIS), moderate (L2: MIA, IA Grade 1), and high (L3: IA Grade 2, IA Grade 3). The goal is to classify lung cancer subtypes into different levels of invasiveness (low, moderate, or high risk).

4.2. Data preprocessing

4.2.1. MIMIC dataset preprocessing

For MIMIC, we utilize electronic health record (EHR) data from the MIMIC-III database, along with chest X-ray images and their corresponding radiology reports from the MIMIC-CXR dataset. The matching of EHR and CXR data is based on 'stay_id', which uniquely identifies specific hospital stays, allowing the association of corresponding X-rays with those stays. Initially, 'subject_id', the unique patient identifier, is used to merge datasets and group records by individual patients. Besides, for REP and CXR image matching, both 'subject_id' and 'study_id' are used to establish a direct correspondence between radiology reports and X-ray images.

For the electronic health record (EHR) modality, we utilize an identical set of 17 clinical variables as employed in prior work [50]. They include five categorical variables and twelve continuous variables. The categorical variables are: capillary refill rate, Glasgow Coma Scale eye opening, Glasgow Coma Scale motor response, Glasgow Coma Scale verbal response, and Glasgow Coma Scale total score. The continuous variables include: diastolic blood pressure, fraction of inspired oxygen, glucose level, heart rate, height, mean blood pressure, oxygen saturation, respiratory rate, systolic blood pressure, temperature, weight, and pH. To ensure consistency and maintain a comparable input across patients, all clinical variables are sampled at two-hour intervals. For categorical features, one-hot encoding is employed, transforming each category into a binary vector representation. Continuous variables are standardized by subtracting the mean and dividing by the standard deviation. After pre-processing and encoding, the EHR modality is represented as a vector of size 76 for each time step.

Regarding the radiology report (REP) modality, the "Findings" sections in these reports provide detailed textual descriptions of the observations made by radiologists. The length of these sections varies

between 10 and 280 words across samples. Through a series of experiments, we determine that truncating or padding the reports to a uniform length of 150 words (99.35% of the reports have fewer than 150 words). For text embedding, we use the `bert-base-uncased` model from BERT [56], utilizing the pre-trained version without task-specific fine-tuning, and each token in the report is converted into a 768-dimensional vector.

For the chest radiographs (CXR), we first resize all images to a fixed size of 224 × 224 pixels to standardize the input across the dataset and replicate each image across three channels. Then, we employ a ResNet34 [57] convolutional neural network to extract relevant features from the resized images. After the average pooling layer of the convolutional network where n = 512, the extracted features are then used as the input representation for the image modality.

4.2.2. Lung cancer pathological subtyping dataset preprocessing

For lung cancer pathological subtyping dataset, it includes clinical textual report (REP), PET and CT imaging sequences.

The clinical textual report (REP) includes patient-specific information such as personal history, past medical history, surgical history, physical examination details, diagnostic basis, and surgical indications. We concatenate all these textual inputs to form a unified report (REP) modality, which serves as a comprehensive summary of the patient's clinical background. The text is encoded using `bert-base-chinese` for effective representation of the Chinese language medical reports. The length of the clinical texts ranges from 996 to 1748 characters, with the 89.27% samples having a length between 1400 and 1600 characters. Given the relatively consistent length, we set the maximum sequence length to 1748 characters and pad shorter texts accordingly, and each token in the report is converted to a 768-dimensional vector.

PET and CT imaging sequences are sliced along the z-axis to generate 2D slices, and each 2D slice is resized to 224 × 224 pixels before being fed into the model. For feature extraction, we utilize the ResNet34 architecture, similar to the approach used for MIMIC-CXR. The PET and CT slices are treated as separate modalities, allowing the model to learn distinct yet complementary features from both types of imaging.

4.3. Hyperparameter settings

The key hyperparameters in our framework include the number of recurrent adapters (RAs) in the Missing Modality Imputation process, the weighting strength for DWFuse (α, β), and the loss weights ($\lambda_1, \lambda_2, \lambda_3$). Specifically, we set the number of RAs to 5, α and β to 0.1 and 0.5, and $\lambda_1, \lambda_2, \lambda_3$ to 1, 10, and 0.5, respectively. The selection process for these hyperparameters begins with leveraging prior work's experience, where we initialize hyperparameters based on values reported in similar studies. Then, we perform cross-validation and conduct a grid search over a range of potential values around the initial settings. Finally, the values that achieve the best AUROC on the respective validation set are selected.

4.4. Comparison method

We evaluate our framework against several baseline and state-of-the-art multi-modal fusion methods, including Concatenate, Sum, SE-Gate, G-Blend, OGM-GE, PMR, and MedFuse. Concatenate involves concatenating the multimodal embeddings directly. Sum refers to summing the multimodal embeddings. SE-Gate is based on Squeeze-and-Excitation Networks, which use an attention mechanism to recalibrate the importance of each modality. G-Blend is Gradient-Blending, which computes an optimal blending of modalities based on their overfitting behaviors. OGM-GE uses on-the-fly gradient modulation to adaptively control the optimization of each modality by monitoring the discrepancy in their contributions towards the learning objective. PMR introduces a prototype-based entropy regularization term during the early training stage to alleviate suppression from the dominant modality and prevent premature convergence. MedFuse is an LSTM-based multi-modal fusion method. To deepen the comparative analysis, we ensure that the hyperparameter settings for each method are configured to the best values reported in their respective original papers. Furthermore, except for the substituted fusion method, all other configurations of our framework remain consistent across experiments to ensure a fair comparison.

Table 2
Dominated uni-modal models consistently outperform the multi-modal models (concatenated fusion) for all three different tasks.

In-hospital mortality				
Multi-modality	AUROC	Dominated uni-modality	AUROC	Drop
EHR+CXR+REP	0.811	EHR	0.828	−0.017
EHR+REP	0.817	EHR	0.828	−0.011
EHR+CXR	0.819	EHR	0.828	−0.009
CXR+REP	0.724	CXR	0.741	−0.017
Phenotyping				
Multi-modality	AUROC	Dominated uni-modality	AUROC	Drop
EHR+CXR+REP	0.696	EHR	0.714	−0.018
EHR+REP	0.694	EHR	0.714	−0.020
EHR+CXR	0.701	EHR	0.714	−0.013
CXR+REP	0.654	CXR	0.664	−0.010
Pathological subtyping				
Multi-modality	AUROC	Dominated uni-modality	AUROC	Drop
CT+PET+REP	0.812	CT	0.833	−0.021
CT+PET	0.815	CT	0.833	−0.018
CT+REP	0.822	CT	0.833	−0.011
PET+REP	0.787	REP	0.803	−0.016

5. Results and analysis

5.1. Modality laziness incurs the failure of multimodal model

In the quest to leverage diverse data types for enhanced model performance through multi-modal networks, an intriguing counterintuitive phenomenon often emerges. Although theoretically, multi-modal networks, which integrate multiple streams of input data, should surpass any uni-modal model due to their richer information set, practical implementations frequently fall short. As shown in Table 2, in the task of in-hospital mortality prediction, integrating multi-modalities including electronic health records (EHR), radiographic images (CXR), and textual clinical reports (REP) into a fused predictive model by concatenating leads to the performance degradation, compared with the dominated uni-modal models. Similar results occur in the phenotyping prediction and pathological subtyping prediction tasks. This phenomenon suggests the presence of "modality laziness", where certain uninformative modalities in the multimodal model do not contribute effectively to predictions, or even exert a negative influence due to noise. For example, in the MIMIC dataset, the EHR data often dominate, overshadowing valuable insights that could potentially be derived from radiographic images and textual reports. This results in a significant underutilization of the multimodal data's potential synergies.

5.2. DWFuse mitigates the multimodal optimization dilemma

The challenge of effectively leveraging the complementary strengths of various modalities in a unified predictive model is addressed by our innovative fusion technique, Dynamic Weighted Fuse (DWFuse). This module aims to address the issue of "modality laziness" by dynamically quantifying the relationships between multiple modalities for specific instances. It leverages the strengths of the more informative modalities, allowing them to exert greater predictive power while minimizing the negative influence of less informative ones.

Table 3
Comparative performances of various fusion methods across different modality combinations for MIMIC-in-hospital mortality prediction task.

Modalities			Fusion methods							
EHR	CXR	REP	Concat	Sum	SE-Gate	G-Blend	OGM-GE	PMR	MedFuse	DWFuse
●	●	○	0.819	0.815	0.822	0.830	0.828	0.830	0.839	0.847*
●	○	●	0.817	0.815	0.820	0.825	0.829	0.831	0.827	0.834*
○	●	●	0.724	0.725	0.727	0.738	0.736	0.734	0.732	0.748*
●	●	●	0.811	0.807	0.815	0.831	0.836	0.834	0.845	0.859*

* Indicates a statistically significant improvement over all other methods ($p < 0.05$), based on a paired t-test conducted using performance scores from 5-fold cross-validation.

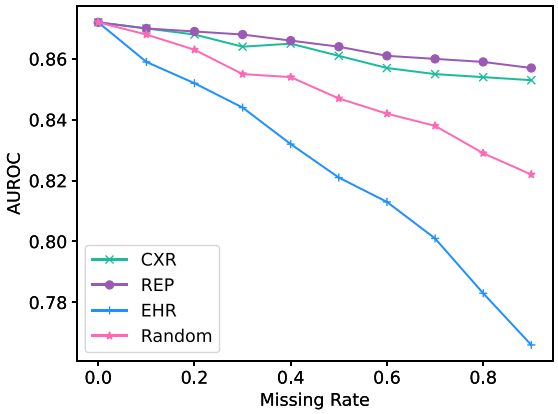

Fig. 4. Impact of increasing missing rate on AUROC across each uni-modality (i.e., CXR, REP, and EHR) or randomly missed any modality (i.e., Random) for MIMIC-in-hospital mortality prediction task.

Our experiments demonstrate that DWFuse outperforms all compared fusion methods across various combinations of multi-modalities. As presented in Table 3, DWFuse consistently achieves superior performance compared to other advanced fusion strategies, including Concatenation [58], Summation [59], SE-Gate [60], G-Blend [28], OGM-GE [29], PMR [10], and MedFuse [50]. In the partly modality combinations of EHR+CXR, EHR+REP, and CXR+REP, DWFuse also shows remarkable improvements, with AUROC scores increasing from 0.819, 0.817, and 0.724 in concatenation to 0.847, 0.834, and 0.748, respectively. Even more impressively, in the full modality combination (i.e., EHR+CXR+REP), DWFuse elevates the AUROC from 0.811 to 0.859, surpassing all the compared methods. To evaluate the robustness of the performance improvements of our proposed DWFuse method over baseline fusion methods, we conduct a statistical significance analysis. Specifically, we employ paired t-tests to compare the results of DWFuse with each of the baseline methods across all modality combinations for the MIMIC in-hospital mortality prediction task. The results of the statistical tests indicate that the improvements achieved by DWFuse are statistically significant compared to all other fusion methods ($p < 0.05$) for each modality combination.

Moreover, DWFuse unlocks the predictive potential of the multi-modal model, surpassing the performance of dominant uni-modal models, i.e., DWFuse advanced multi-modal performance of 0.859 better than dominated EHR uni-modal performance of 0.828. These results underscore the effectiveness of DWFuse in mitigating the multi-modal optimization dilemma. By adaptively weighting the input from each modality according to its predictive power and contextual relevance, DWFuse ensures that the integration of multimodal data is not only more balanced but also more synergistic.

5.3. Enhancing modality integration with missing data imputation

One significant challenge in multi-modal models is handling inconsistencies and gaps within the data—a problem often exacerbated by the missing data inherent in real-world clinical settings. This section explores the effectiveness of our proposed missing data imputation (IMP) module, which infers the embeddings of missing modalities from those of available one based on the underlying connections between modalities.

As shown in Fig. 4, we first explore the impact of different missing rate on model performance across each uni-modality (i.e., CXR, REP, and EHR) or randomly missed any modality (i.e., Random) on the in-hospital mortality prediction task. The 'Random' line indicates the AUROC when one of these modalities is randomly missing. As seen, the performance decline of the model vary widely and drastically, with EHR modality showing the most significant drop, highlighting its critical role in the in-hospital mortality prediction task. Moreover, the model's performance decline varies across different modalities, underscoring the need for tailored data imputation strategies for different patient instances and modalities. As observed, there are two key findings: (1) The performance decline is significant and drastic, with the EHR modality showing the most substantial drop. This highlights the critical role of the EHR modality in the in-hospital mortality prediction task; (2) The degree of performance decline varies across different modalities, underscoring the need for tailored data imputation strategies for different patient instances and modalities.

Then, we verify the effectiveness of our proposed data imputation method by experimenting (1) across various modality combinations, and (2) across different missing rates in Figs. 5 and 6, respectively. The results demonstrate that the introduction of the imputation strategy significantly improves performance. In detail, we compare the model's performance with and without the data imputation module (Imp and No-imp) applied across different modality combinations, including CXR + Rep, CXR + EHR, Rep + EHR, and CXR + Rep + EHR, against increasing missing data rates from 0% to 90%. Models implementing imputation start with robust AUROC values and exhibit a slower decline in performance as the missing data rate increases, compared to the steeper performance drop observed in models without imputation. For instance, in the combination of all three modalities (i.e., CXR + Rep + EHR), the imputation model starts near an AUROC of 0.87 and remains above 0.86 even at 90% missing data, significantly outperforming the non-imputation model which declines to approximately 0.83. This evidence highlights the importance of incorporating effective imputation methods in clinical predictive analytics, enhancing the reliability and accuracy of predictions crucial for informed decision-making in healthcare settings, particularly in scenarios with substantial data incompleteness.

Furthermore, Fig. 7 illustrates the evolution of missing data imputation effectiveness during the training process for the pathological subtyping prediction task. The red curve represents the similarity matrix between the imputed missing data and the ground truth missing data, while the blue curve represents the AUROC. It is evident that the imputed data increasingly resembles the ground truth values as training progresses, demonstrating the improving imputation accuracy of our method. Also, Fig. 8 shows the T-SNE visualization of the original and imputed uni-modality embeddings. For example, EHR denotes the original uni-modality embeddings of electronic health records, while EHR-imputed represents the imputed ones. This visualization highlights the preservation of data structure post-imputation, illustrating how closely imputed embeddings align with their original counterparts in the embedded space.

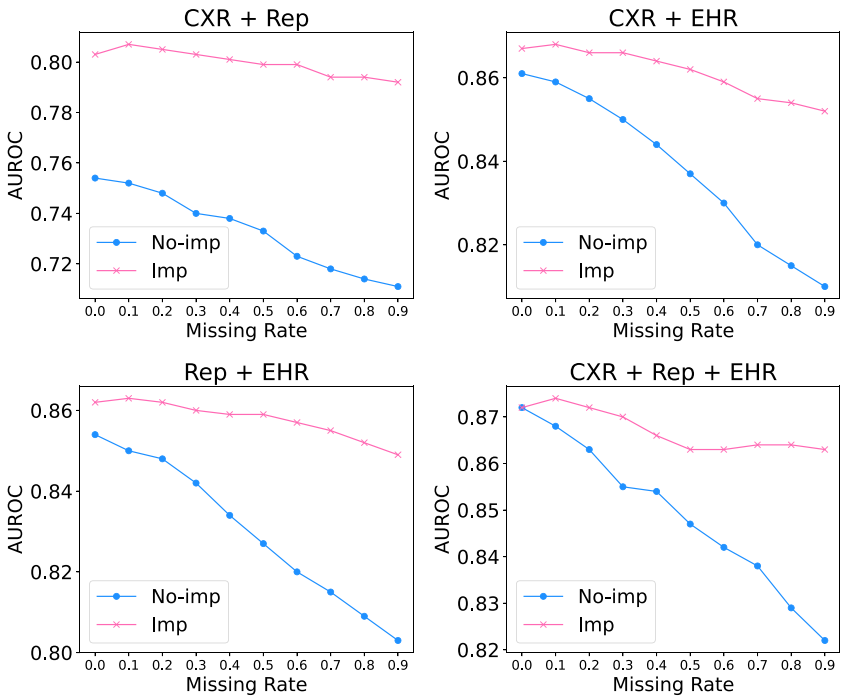

Fig. 5. Impact of data imputation on AUROC *across various modality combinations* for the MIMIC in-hospital mortality prediction task. Notably, even when the missing rate is zero for the two given modalities, our method can still impute the third modality and achieve improved performance.

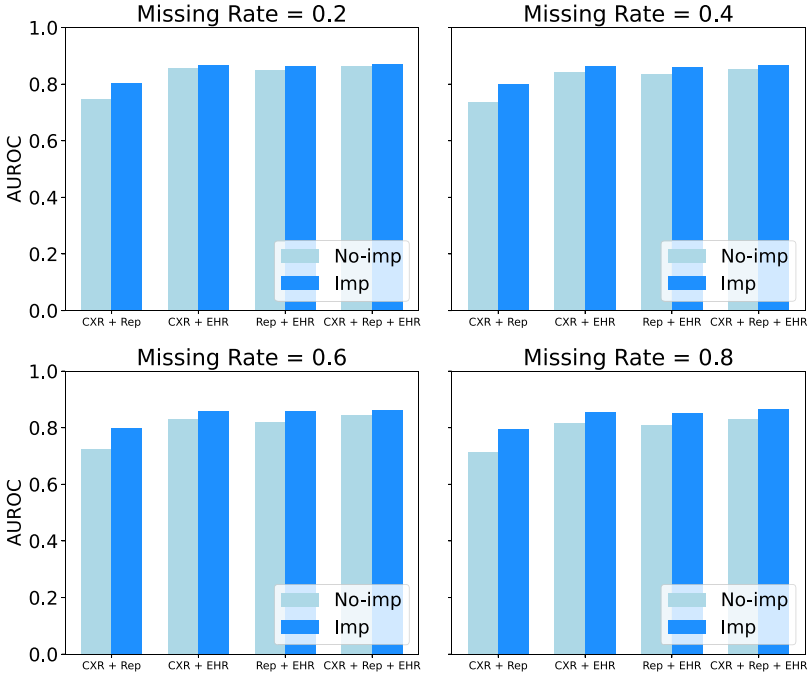

Fig. 6. Impact of data imputation on AUROC *across different missing rates* for MIMIC-in-hospital mortality prediction task.

5.4. Ablation study of designed components

Table 4 lists the impact of the key components in our proposed framework across various modality combinations. Initially, the row (a) of each task represents the baseline, which removes all proposed components and fuses multi-modalities in a basic concatenation manner. Based on this, the introduction of DWFuse alone significantly improves predictive outcomes by dynamically weighting each modality's input, from 0.811 to 0.859 for in-hospital mortality prediction and from 0.812 to 0.868 for pathological subtyping prediction. In row (c), adding TLA training strategy further boosts performance by enabling the traceability and activation of the lazy modalities combinations, which is evident in the incremental AUROC improvements—for instance, increasing from 0.834 to 0.854 in REP+EHR modality combination for in-hospital mortality prediction. Additionally, introducing the IMP-forward module in row (d) improves model performance, and the inclusion of IMP-backward further enhances performance, as shown in

Table 4
Ablation performances with different designed components.

In-hospital mortality								
Components					CXR+REP	CXR+EHR	REP+EHR	CXR+REP+EHR
	DWFuse	TLA	IMP-fw	IMP-bw				
(a)					0.724	0.819	0.817	0.811
(b)	✓				0.748	0.847	0.834	0.859
(c)	✓	✓			0.754	0.861	0.854	0.872
(d)	✓	✓	✓		0.772	0.863	0.862	0.870
(e)	✓		✓	✓	0.789	0.855	0.854	0.863
(f)	✓	✓	✓	✓	0.803	0.867	0.862	0.872
Pathological subtyping								
Components					CT+PET	CT+REP	PET+REP	CT+PET+REP
	DWFuse	TLA	IMP-fw	IMP-bw				
(a)					0.815	0.822	0.787	0.812
(b)	✓				0.845	0.849	0.819	0.868
(c)	✓	✓			0.859	0.861	0.833	0.889
(d)	✓	✓	✓		0.863	0.869	0.842	0.890
(e)	✓		✓	✓	0.860	0.863	0.850	0.868
(f)	✓	✓	✓	✓	0.869	0.874	0.859	0.894

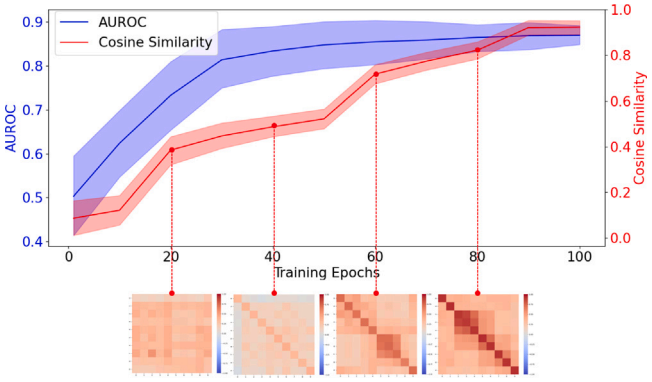

Fig. 7. Evolution of missing data imputation effectiveness during training progress for pathological subtyping prediction task. The red curve illustrates the similarity matrix between imputed missing data and ground truth missing data. The blue curve represents the AUROC. It can be seen that the imputed data increasingly resembles ground truth values with the progression of training, demonstrating the improving imputation accuracy of our method.

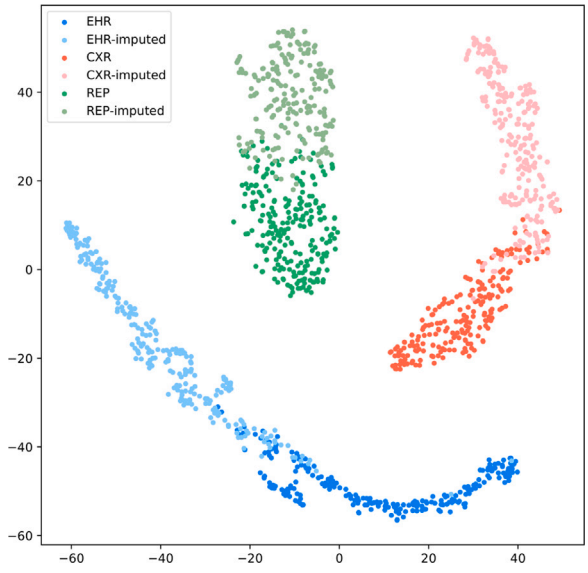

Fig. 8. The T-SNE visualization of the original ground truth embeddings and imputed embeddings of each uni-modality for MIMIC-phenotype classification task.

row (f). For instance, the inclusion of backward imputation improves the CXR+REP performance from 0.772 to 0.803 in the In-hospital Mortality task and increases the PET+REP performance from 0.842 to 0.859 in Pathological Subtyping. These results show that backward imputation enhances the consistency and accuracy of reconstructed modalities, ultimately leading to better overall predictive performance. Moreover, the complete integration of DWFuse, TLA, and IMP yields the most substantial gains, with AUROC reaching up to 0.872 in the CXR+REP+EHR combination for in-hospital mortality prediction, and up to 0.894 for pathological subtyping prediction. It can be seen that our framework not only mitigates the negative impact of any underperforming lazy modality but also unleashes its potential predictive power, leading to more accurate predictions. Such strategies are crucial in clinical settings where low-quality data is prevalent, while precise multimodal data interpretation can significantly influence patient outcomes.

6. Conclusion and future work

In this paper, we introduce a novel framework designed to address the inherent challenges in multimodal learning, particularly within the medical domain with low-quality data. Traditional methods often struggle with incomplete or missing modalities, which significantly impact prediction accuracy and robustness. Our framework effectively handles any combination of missing modalities through autonomous imputation and employs a dynamic weighted fusion technique that explicitly quantifies and leverages inter-modality relationships. Additionally, our approach identifies and activates "lazy modalities", thereby enhancing overall prediction accuracy and reliability. By dynamically adapting to changing data scenarios and providing reliable, traceable predictive outputs, our framework represents a significant advancement in the field of multimodal learning. Experimental validation on MIMIC-III, MIMIC-IV, and a proprietary lung cancer pathological subtyping dataset, demonstrates the robustness and versatility of our framework.

A notable design choice in this work is the use of BERT as a fixed text encoder. Fine-tuning BERT led to poorer results, so we opt to use it without task-specific adaptation. However, future research could explore partially fine-tuning BERT or utilizing domain-specific models, such as Medical BERT, to enhance medical knowledge representation. While the primary objective of this work is to propose a generalizable and scalable framework, incorporating specialized text encoding methods may further improve accuracy in specific medical applications.

Future work will also focus on enhancing the scalability and computational efficiency of our framework, making it suitable for large-scale, real-time applications. We aim to extend the framework's capabilities to handle a wider range of data modalities and test its applicability

across other domains, such as autonomous systems and multimedia processing.

Additionally, we plan to integrate our framework into clinical workflows to enable real-time support and decision-making for healthcare professionals. This will include developing user-friendly interfaces and visualization tools to allow end-users to interpret predictions and understand the contributions of different modalities. Ensuring data security and privacy, particularly in medical applications, will also be a priority, as compliance with relevant regulations is essential to safeguard sensitive information. By pursuing these directions, we aim to enhance the practical utility of our framework and promote the broader adoption of robust multimodal learning systems across diverse fields.

CRediT authorship contribution statement

Yixuan Wu: Writing – original draft, Visualization, Software, Methodology, Investigation. **Jintai Chen:** Writing – review & editing, Validation, Supervision. **Lianting Hu:** Formal analysis, Data curation. **Hongxia Xu:** Visualization, Supervision. **Huiying Liang:** Writing – review & editing, Resources, Project administration, Investigation, Conceptualization. **Jian Wu:** Supervision, Resources, Project administration, Funding acquisition.

Funding

This research was partially supported by National Natural Science Foundation of China under grants No.82202984, No. 62076076, No. 82122036, and No. 12326612, Basic and Applied Basic Research Foundation of Guangdong Province, China under grant No. 2022A1515110722, and No. 2024A1515011750.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

[1] J. Liu, H. Zheng, D. Huo, Y. Hao, D. Niyato, S.A. Alqahtani, M. Chen, Big fiber slicing for dynamic multi-modal multi-preference applications of smart fabrics, IEEE Internet Things J. (2024).

[2] Y. Wu, J. Chen, J. Yan, Y. Zhu, D.Z. Chen, J. Wu, GCL: Gradient-guided contrastive learning for medical image segmentation with multi-perspective meta labels, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 463–471.

[3] C. Gan, Y. Tang, X. Fu, Q. Zhu, D.K. Jain, S. García, Video multimodal sentiment analysis using cross-modal feature translation and dynamical propagation, Knowl.-Based Syst. (2024) 111982.

[4] Y. Wu, Z. Zhang, C. Xie, F. Zhu, R. Zhao, Advancing referring expression segmentation beyond single image, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 2628–2638.

[5] Y. Wang, Y. Wu, S. Tang, W. He, X. Guo, F. Zhu, L. Bai, R. Zhao, J. Wu, T. He, et al., Hulk: A universal knowledge translator for human-centric tasks, 2023, arXiv preprint arXiv:2312.01697.

[6] J. Yan, J. Chen, Y. Wu, D.Z. Chen, J. Wu, T2g-former: organizing tabular features into relation graphs promotes heterogeneous feature interaction, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 10720–10728.

[7] J. Wang, T. Chen, J. Chen, Y. Wu, Y. Xu, D. Chen, H. Ying, J. Wu, PoCo: A self-supervised approach via polar transformation based progressive contrastive learning for ophthalmic disease diagnosis, 2024, arXiv preprint arXiv:2403.19124.

[8] Y. Wu, B. Zheng, J. Chen, D.Z. Chen, J. Wu, Self-learning and one-shot learning based single-slice annotation for 3d medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 244–254.

[9] J. Wen, Z. Zhang, L. Fei, B. Zhang, Y. Xu, Z. Zhang, J. Li, A survey on incomplete multiview clustering, IEEE Trans. Syst. Man Cybern.: Syst. 53 (2) (2022) 1136–1149.

[10] Y. Fan, W. Xu, H. Wang, J. Wang, S. Guo, Pmr: Prototypical modal rebalance for multimodal learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 20029–20038.

[11] M. Salvi, H.W. Loh, S. Seoni, P.D. Barua, S. García, F. Molinari, U.R. Acharya, Multi-modality approaches for medical support systems: A systematic review of the last decade, Inf. Fusion (2023) 102134.

[12] H. Feng, Q. Li, W. Wang, A.K. Bashir, A.K. Singh, J. Xu, K. Fang, Security of target recognition for UAV forestry remote sensing based on multi-source data fusion transformer framework, Inf. Fusion 112 (2024) 102555.

[13] A.E. Johnson, T.J. Pollard, L. Shen, L.-w.H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Sci. Data 3 (1) (2016) 1–9.

[14] A.E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T.J. Pollard, S. Hao, B. Moody, B. Gow, et al., MIMIC-IV, a freely accessible electronic health record dataset, Sci. Data 10 (1) (2023) 1.

[15] D. Zhang, D. Shen, A.D.N. Initiative, et al., Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease, Neuroimage 59 (2) (2012) 895–907.

[16] F. Liu, C.-Y. Wee, H. Chen, D. Shen, Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's Disease and mild cognitive impairment identification, Neuroimage 84 (2014) 466–475.

[17] K.-H. Thung, C.-Y. Wee, P.-T. Yap, D. Shen, A.D.N. Initiative, et al., Neurodegenerative disease diagnosis using incomplete multi-modality data via matrix shrinkage and completion, Neuroimage 91 (2014) 386–400.

[18] W. van Loon, M. Fokkema, F. de Vos, M. Koini, R. Schmidt, M. de Rooij, Imputation of missing values in multi-view data, Inf. Fusion (2024) 102524.

[19] Q. Zhou, T. Chen, H. Zou, X. Xiao, Uncertainty-aware incomplete multimodal fusion for few-shot Central Retinal Artery Occlusion classification, Inf. Fusion 104 (2024) 102200.

[20] W. Shao, L. He, P.S. Yu, Multiple incomplete views clustering via weighted nonnegative matrix factorization with regularization, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2015, pp. 318–334.

[21] H. Zhao, H. Liu, Y. Fu, Incomplete multi-modal visual data grouping, in: IJCAI, 2016, pp. 2392–2398.

[22] Y. Ye, X. Liu, Q. Liu, J. Yin, Consensus kernel K-means clustering for incomplete multiview data, Comput. Intell. Neurosci. 2017 (1) (2017) 3961718.

[23] X. Liu, Incomplete multiple kernel alignment maximization for clustering, IEEE Trans. Pattern Anal. Mach. Intell. 46 (3) (2021) 1412–1424.

[24] J. Wen, Z. Zhang, Z. Zhang, Z. Wu, L. Fei, Y. Xu, B. Zhang, Dimc-net: Deep incomplete multi-view clustering network, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 3753–3761.

[25] J. Wen, Z. Wu, Z. Zhang, L. Fei, B. Zhang, Y. Xu, Structural deep incomplete multi-view clustering network, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 3538–3542.

[26] C. Liu, J. Wen, X. Luo, C. Huang, Z. Wu, Y. Xu, Dicnet: Deep instance-level contrastive network for double incomplete multi-view multi-label classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 8807–8815.

[27] C. Zhang, Z. Han, H. Fu, J.T. Zhou, Q. Hu, et al., CPM-Nets: Cross partial multi-view networks, Adv. Neural Inf. Process. Syst. 32 (2019).

[28] W. Wang, D. Tran, M. Feiszli, What makes training multi-modal classification networks hard? in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12695–12705.

[29] X. Peng, Y. Wei, A. Deng, D. Wang, D. Hu, Balanced multimodal learning via on-the-fly gradient modulation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 8238–8247.

[30] N. Wu, S. Jastrzebski, K. Cho, K.J. Geras, Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks, in: International Conference on Machine Learning, PMLR, 2022, pp. 24043–24055.

[31] G. Muhammad, F. Alshehri, F. Karray, A. El Saddik, M. Alsulaiman, T.H. Falk, A comprehensive survey on multimodal medical signals fusion for smart healthcare systems, Inf. Fusion 76 (2021) 355–375.

[32] Y. Sun, S. Mai, H. Hu, Learning to balance the learning rates between various modalities via adaptive tracking factor, IEEE Signal Process. Lett. 28 (2021) 1650–1654.

[33] Y. Zhou, X. Liang, S. Zheng, H. Xuan, T. Kumada, Adaptive mask co-optimization for modal dependence in multimodal learning, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2023, pp. 1–5.

[34] Y. Wei, R. Feng, Z. Wang, D. Hu, Enhancing multimodal cooperation via sample-level modality valuation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 27338–27347.

[35] S.C. Kulkarni, P.P. Rege, Pixel level fusion techniques for SAR and optical images: A review, Inf. Fusion 59 (2020) 13–29.

[36] X. Cheng, Y. Zhong, Y. Dai, P. Ji, H. Li, Noise-aware unsupervised deep lidar-stereo fusion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6339–6348.

[37] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, C.-L. Tai, Transfusion: Robust lidar-camera fusion for 3d object detection with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1090–1099.
[38] B. Rajalingam, F. Al-Turjman, R. Santhoshkumar, M. Rajesh, Intelligent multimodal medical image fusion with deep guided filtering, Multimedia Syst. 28 (4) (2022) 1449–1463.
[39] Q. Guihong, Z. Dali, Y. Pingfan, Medical image fusion by wavelet transform modulus maxima, Opt. Express 9 (4) (2001) 184–190.
[40] A. Achim, C. Canagarajah, D. Bull, Complex wavelet domain image fusion based on fractional lower order moments, in: 2005 7th International Conference on Information Fusion, Vol. 1, IEEE, 2005, pp. 7–pp.
[41] L. Gjesteby, B. De Man, Y. Jin, H. Paganetti, J. Verburg, D. Giantsoudi, G. Wang, Metal artifact reduction in CT: where are we after four decades? Ieee Access 4 (2016) 5826–5849.
[42] S. Changpinyo, P. Sharma, N. Ding, R. Soricut, Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3558–3568.
[43] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, Z. Liu, Weakly aligned cross-modal learning for multispectral pedestrian detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5127–5137.
[44] P. Sharma, N. Ding, S. Goodman, R. Soricut, Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2556–2565.
[45] F. Radenovic, A. Dubey, A. Kadian, T. Mihaylov, S. Vandenhende, Y. Patel, Y. Wen, V. Ramanathan, D. Mahajan, Filtering, distillation, and hard negatives for vision-language pre-training, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 6967–6977.
[46] S.Y. Gadre, G. Ilharco, A. Fang, J. Hayase, G. Smyrnis, T. Nguyen, R. Marten, M. Wortsman, D. Ghosh, J. Zhang, et al., Datacomp: In search of the next generation of multimodal datasets, Adv. Neural Inf. Process. Syst. 36 (2024).
[47] S. Wang, M.B. McDermott, G. Chauhan, M. Ghassemi, M.C. Hughes, T. Naumann, Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii, in: Proceedings of the ACM Conference on Health, Inference, and Learning, 2020, pp. 222–235.
[48] M. Sadeghi, X. Alameda-Pineda, Switching variational auto-encoders for noise-agnostic audio-visual speech enhancement, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2021, pp. 6663–6667.
[49] Z. Li, Y. Jiang, M. Lu, R. Li, Y. Xia, Survival prediction via hierarchical multimodal co-attention transformer: A computational histology-radiology solution, IEEE Trans. Med. Imaging 42 (9) (2023) 2678–2689.
[50] N. Hayat, K.J. Geras, F.E. Shamout, MedFuse: Multi-modal fusion with clinical time-series data and chest X-ray images, in: Machine Learning for Healthcare Conference, PMLR, 2022, pp. 479–503.
[51] M. Wang, S. Fan, Y. Li, H. Chen, Missing-modality enabled multi-modal fusion architecture for medical data, 2023, arXiv preprint arXiv:2309.15529.
[52] F.E. Shamout, Y. Shen, N. Wu, A. Kaku, J. Park, T. Makino, S. Jastrzębski, J. Witowski, D. Wang, B. Zhang, et al., An artificial intelligence system for predicting the deterioration of COVID-19 patients in the emergency department, NPJ Digit. Med. 4 (1) (2021) 80.
[53] W. Shao, T. Wang, L. Sun, T. Dong, Z. Han, Z. Huang, J. Zhang, D. Zhang, K. Huang, Multi-task multi-modal learning for joint diagnosis and prognosis of human cancers, Med. Image Anal. 65 (2020) 101795.
[54] D. Ho, I.B.H. Tan, M. Motani, Predictive models for colorectal cancer recurrence using multi-modal healthcare data, in: Proceedings of the Conference on Health, Inference, and Learning, 2021, pp. 204–213.
[55] L.A. Vale-Silva, K. Rohr, Long-term cancer survival prediction using multimodal deep learning, Sci. Rep. 11 (1) (2021) 13505.
[56] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.
[57] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[58] J. Duan, J. Xiong, Y. Li, W. Ding, Deep learning based multimodal biomedical data fusion: An overview and comparative review, Inf. Fusion (2024) 102536.
[59] T. Shaik, X. Tao, L. Li, H. Xie, J.D. Velásquez, A survey of multimodal information fusion for smart healthcare: Mapping the journey from data to wisdom, Inf. Fusion (2023) 102040.
[60] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.