DrFuse: Learning Disentangled Representation for Clinical Multi-Modal Fusion with Missing Modality and Modal Inconsistency

Wenfang Yao[*1], **Kejing Yin**[*2], **William K. Cheung**[2], **Jia Liu**[3], **Jing Qin**[1]

[1]School of Nursing, The Hong Kong Polytechnic University
[2]Department of Computer Science, Hong Kong Baptist University
[3]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
dorothy.yao@polyu.edu.hk, {cskjyin, william}@comp.hkbu.edu.hk, jia.liu@siat.ac.cn, harry.qin@polyu.edu.hk

Abstract

The combination of electronic health records (EHR) and medical images is crucial for clinicians in making diagnoses and forecasting prognoses. Strategically fusing these two data modalities has great potential to improve the accuracy of machine learning models in clinical prediction tasks. However, the asynchronous and complementary nature of EHR and medical images presents unique challenges. Missing modalities due to clinical and administrative factors are inevitable in practice, and the significance of each data modality varies depending on the patient and the prediction target, resulting in inconsistent predictions and suboptimal model performance. To address these challenges, we propose `DrFuse` to achieve effective clinical multi-modal fusion. It tackles the missing modality issue by disentangling the features shared across modalities and those unique within each modality. Furthermore, we address the modal inconsistency issue via a disease-wise attention layer that produces the patient- and disease-wise weighting for each modality to make the final prediction. We validate the proposed method using real-world large-scale datasets, MIMIC-IV and MIMIC-CXR. Experimental results show that the proposed method significantly outperforms the state-of-the-art models.

Introduction

Clinicians rely on data from various sources, including electronic health records (EHR) and medical imaging, to make diagnoses and forecast prognoses (Aljondi and Alghamdi 2020). For instance, when diagnosing pneumonia, EHR data like blood tests provides information about the patient's infection status and immune response, while medical images like Chest X-ray (CXR) can reveal the extent of inflammation in the lungs (Hoare and Lim 2006). Integrating these data modalities could shed light on a more comprehensive and accurate understanding of the patient's health condition, potentially leading to a better clinical outcome (Huang et al. 2020a). With the increasing availability of digital clinical data, research efforts have recently been made to employ multi-modal machine learning approaches to improve the performance of clinical prediction tasks, including disease prediction (Hayat, Geras, and Shamout 2022) and mortality prediction (Lin et al. 2021).

---

[*]These authors contributed equally.

The multi-modal data fusion, *i.e.*, the process of combining different data modalities, plays a central role in the effective utilization of multi-modal clinical data. Despite the recent effort, their applications to real-world data are still hindered due to the complex and complementary nature of multi-modal clinical data. Specifically, there are fundamentally challenging issues that need to be addressed:

Challenge 1: Missing modality in a highly heterogeneous setting. Many existing work on clinical multi-modal learning assumes that both EHR and medical images are available for all training and testing samples, which is not practical in real-world clinical settings. For instance, the MIMIC-IV (Johnson et al. 2023), a real-world ICU dataset, has less than 20% of patients with X-ray images. Many in-hospital patients requiring X-ray scans cannot undergo the procedure due to clinical or administrative reasons, resulting in a significant number of patients with missing modalities (Huang et al. 2020a). Similar problems exist in other domains like tumor segmentation on multi-modal MRI images (Zhao, Yang, and Sun 2022), where generative machine learning models are commonly used to synthesize the missing modality (Sharma and Hamarneh 2019). However, accurately generating missing medical images using EHR data is infeasible because EHR contains information about a patient's clinical conditions, medical history, and treatments, but they do not provide a detailed enough picture of the patient's anatomy to generate a missing modality of medical imaging, such as a chest X-ray. Late fusion is a common approach to tackling missing modalities in the fusion of EHR and medical imaging, where separate prediction models are learned for different modalities and the fusion happens only in the decision level (Huang et al. 2020a). This approach fails to fully utilize the interactions between modalities, leading to undesirable suboptimal performance. Therefore, effectively capturing the complex interactions between highly heterogeneous modalities while handling missing modalities remains an open challenge.

Challenge 2: Modal inconsistency and patient-specific modal significance. Even with fully observed data modalities, inconsistencies can arise when different modalities, such as EHR and CXR, provide inconsistent or even contradictory information regarding the prediction targets. For example, in mortality prediction, patients with meningitis may be identified as having a high risk of mortality based on EHR

data due to the severity of their symptoms, while their CXR may not show any signs of complications (Brouwer, Tunkel, and van de Beek 2010). Conversely, for patients with pneumothorax, CXR may predict a high risk of mortality while EHR may not indicate so due to the non-specific nature of the symptoms (Zarogoulidis et al. 2014). The patient variation makes it even more challenging as the significance of different data modalities depends on the patient's medical condition. For example, diabetic patients without specific symptoms or conditions are usually not recommended to take X-rays, while those who develop complications like foot or dental problems need X-rays to assist in diagnosing and treatment planning (Ahmad 2016). Without appropriately accounting for such inconsistency and patient-specific significance between modalities, the accuracy of model prediction could be greatly compromised, leading to suboptimal clinical outcomes. How to effectively handle modal inconsistency and patient-dependent modal significance in multimodal learning remains an unresolved research problem.

To address the above challenges, we propose a novel method: *Learning Disentangled Representation for Clinical Multi-Modal Fusion* (`DrFuse`). We hypothesize that EHR and medical images share a common information component. To leverage this shared information, our core idea is to disentangle the shared information from the modality-distinct information of EHR and medical images. By doing so, we learn a shared representation that captures the common information across both modalities, which enables us to make more accurate predictions even when one modality is unavailable, as the shared information can be inferred from the available modality. To further utilize distinct information from each modality and allow the patient-dependent modal significance to be captured, we propose a disease-aware attention fusion module that is regulated by a novel attention weight ranking loss.

To summarize, our main contributions are three-fold:

- We propose `DrFuse` to fully utilize information shared across modalities with disentangled representation learning. It tackles the missing modality issue as the shared information is still preserved with the available modality robustly under an end-to-end learning paradigm.

- `DrFuse` captures the patient-specific significance of EHR and medical images for each prediction target and therefore tackles the modal inconsistency problem. To the best of our knowledge, this is the first work addressing the modal inconsistency issue for highly heterogeneous clinical multi-modal data.

- Our experimental results show that `DrFuse` significantly outperforms state-of-the-art models on the phenotype classification task in the real-world large-scale MIMIC-IV dataset.

Related Work

Multi-modal learning for healthcare. It has been shown that fusing multiple modalities has great potential to enhance machine learning models for clinical tasks such as prognosis prediction (Kline et al. 2022), phenotyping classification (Hayat, Geras, and Shamout 2022) and medical im-

age segmentation (Huang et al. 2020b). Various data modalities, including electronic health records (EHR), clinical notes, Electrocardiogram (ECG), omics, chest X-rays, Magnetic Resonance Imaging (MRI), and computed tomography (CT), have been studied in the context of multi-modal learning (Venugopalan et al. 2021; Mohsen et al. 2022). For example, (Pölsterl, Wolf, and Wachinger 2021) combined 3D image and tabular information for diagnosis. Both (Huang et al. 2020b) and (Zhi et al. 2022) fused CT images and EHR for Pulmonary Embolism(PE) diagnosis.

Missing modality. Although the available modalities are abundant, in practice, some modalities are inevitably missing (Huang et al. 2020a). Late fusion is a common solution to handle the missing modality (Yoo et al. 2019). It aggregates the predictions from each modality with a weighted sum or major voting. As each modality is modeled independently, the interaction across modalities cannot be fully captured and utilized (Huang et al. 2020a). Some recent research adopted generative methods to impute or reconstruct the missing modality on an instance or embedding level for compensation. (Ma et al. 2021) reconstructs the features of missing modality by a Bayesian meta-learning framework. (Hayat, Geras, and Shamout 2022) utilized an LSTM layer to generate a representative vector for general cases. (Zhang et al. 2022) proposed to impute in the latent space with auxiliary information. These methods either require prior knowledge or assume different modalities to be similar. It has also been speculated that results relying on generating missing representation may not be robust (Li et al. 2023). Another method is to disentangle the shared and complementary information across modalities and use the shared information for reconstruction or downstream tasks (Chen et al. 2019; Shen and Gao 2019; Wang et al. 2023). Nevertheless, most of these works focus on modalities with much shared information in common, for example, using four modalities of MRI for brain tumor segmentation. How to handle missing modality in a highly heterogeneous setting, like the fusion of EHR and medical image, remains an open challenge.

Modal inconsistency. The issue of model inconsistency has been recognized in different domains. For example, recent works utilize the inconsistency between image and text to detect fake news (Xiong et al. 2023; Sun et al. 2023). The modal inconsistency issue has also been investigated in sentiment analysis using text and images. However, it has not yet been discussed and addressed in the context of clinical multi-modal learning.

DrFuse: The Proposed Method

Notations

In this work, we focus on making clinical predictions using two modalities: electronic health records (EHR), which are recorded in the form of time series, and chest X-ray images (CXR). We denote the EHR data of the nth patient by $\mathbf{X}^{\text{EHR}}_{(n)} \in \mathbb{R}^{T_n \times J}$, where T_n and J are the length of the time series and the number of features, respectively. We denote the CXR data by $\mathbf{X}^{\text{CXR}}_{(n)}$ and the prediction labels by $\mathbf{y}_n$. The data of patients who have both modalities is denoted by $\mathcal{D}_{\text{paired}} = \{(\mathbf{X}^{\text{EHR}}_{(n)}, \mathbf{X}^{\text{CXR}}_{(n)}, \mathbf{y}_n)\}_{n=1}^{N}$. In

(a) The architecture overview of `DrFuse`.

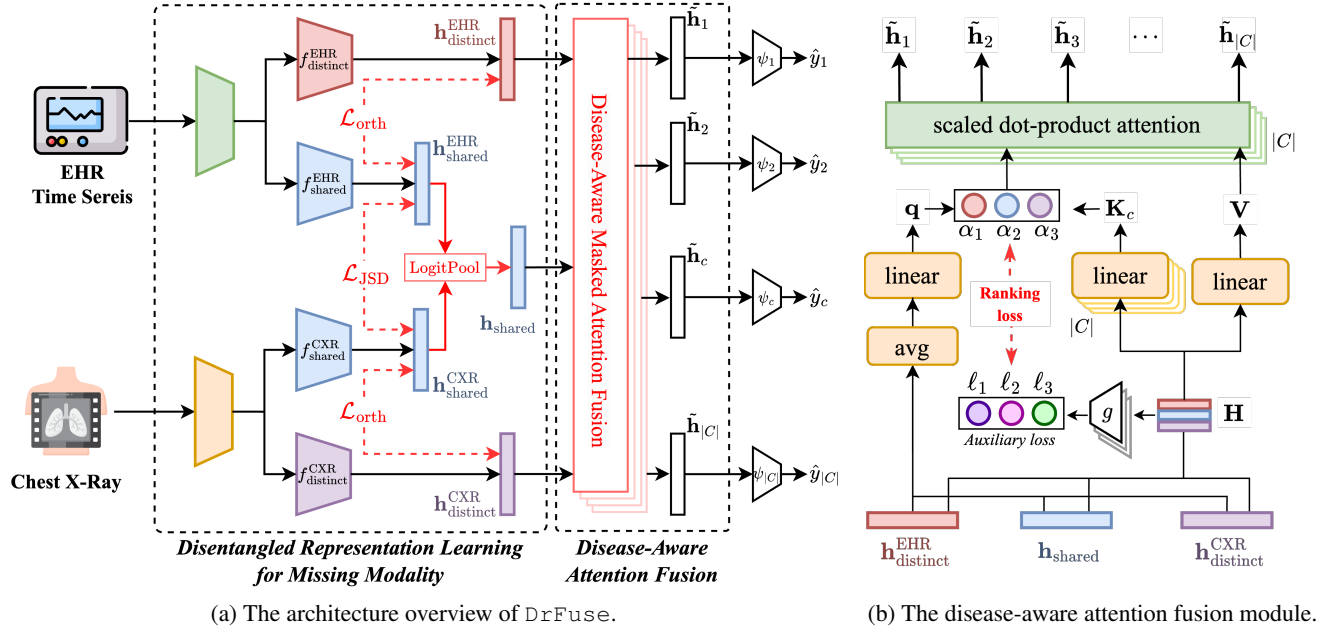(b) The disease-aware attention fusion module.

Figure 1: The overview of the proposed model, `DrFuse`. It consists of two major components. Subfigure (a): A shared representation and a distinct representation are learned from EHR and CXR, where the shared ones are aligned by minimizing the Jensen–Shannon divergence (JSD). A novel logit pooling is proposed to fuse the shared representations. Subfigure (b): The *disease-aware attention fusion* module captures the patient-specific modal significance for different prediction targets by minimizing a ranking loss.

practice, EHRs are routinely recorded in the clinical process but CXR may not always be available. The data of patients who have only EHR data are denoted by $\mathcal{D}_{\text{partial}} = \{(\mathbf{X}_{n'}^{\text{EHR}}, \mathbf{X}_{n'}^{\text{CXR}} = \emptyset, \mathbf{y}_{n'})\}_{n'=1}^{N'}$. To take full advantage of the available data, we use the joint of them as the full dataset, *i.e.*, $\mathcal{D} = \mathcal{D}_{\text{paired}} \cup \mathcal{D}_{\text{partial}}$. To ease the notation, we omit the index of patient n when doing so does not cause confusion.

Overview

An overview of the proposed method is depicted in Fig. 1. It consists of two main components. The *disentangled representation learning* takes the EHR and CXR data as input and generates three representations, the EHR distinct representation $\mathbf{h}_{\text{distinct}}^{\text{EHR}}$, the CXR distinct representation $\mathbf{h}_{\text{distinct}}^{\text{CXR}}$. A novel logit pooling is proposed to generate the cross-modal shared representation $\mathbf{h}_{\text{shared}}$ while achieving effective distribution alignment between the two shared representations. To address the modal inconsistency issue, we propose a *disease-aware attention-based fusion* that adaptively fuses the representations extracted in a patient- and disease-specific manner, where the modal significance for each prediction target can be respected. Finally, the *channel-wise prediction* component makes predictions using the fused representation.

Disentangled Representation Learning

Modal-specific encoders. EHR and CXR are two highly heterogeneous modalities, requiring separate models to encode the raw input data. For each modality, we employ two encoders with the same architecture to extract the shared and distinct representations with dimensions of d.

For EHR data, we use Transformer models (Vaswani et al. 2017) as the encoder, given by:

$$f^{\text{EHR}}(\mathbf{X}) = \text{Transformer}\left([\phi(\mathbf{x}_1) + \delta_1, \ldots, \phi(\mathbf{x}_T) + \delta_T]\right),$$

where $\phi(\mathbf{x}_t)$ projects the raw EHR time series into an embedding space at time step t and δ_t is the positional encoding. To reduce the number of parameters to be learned, we share the first layer in the two Transformer encoders which are expected to extract low-level features. To extract representations from CXR, we use ResNet50 (He et al. 2016) as the encoders for CXR data.

Shared representation alignment and logit pooling. The purpose of learning disentangled representation is to extract common information that is shared across modalities so that this shared information can still be fully utilized even when one modality is missing. To this end, we need to align the distributions of the shared representations generated from EHR and CXR data. We interpret the shared representations $\mathbf{h}_{\text{distinct}}^{\text{CXR}}$ and $\mathbf{h}_{\text{distinct}}^{\text{EHR}}$ as logits of two probability distributions of a latent multivariate binary random variable, and minimize the Jensen–Shannon divergence (JSD) between the induced distributions $P = \sigma(\mathbf{h}_{\text{distinct}}^{\text{EHR}})$ and $Q = \sigma(\mathbf{h}_{\text{distinct}}^{\text{CXR}})$, where $\sigma(\cdot)$ denotes the standard logistic function, mapping the real-value logits $\mathbf{h}$ to a probability value.

The JSD, also known as total divergence to the average, measures the average information that each sample reveals about the source of the distribution from which it is sampled.

Recent work has shown that JSD is more stable, consistent, and insensitive across a diverse range of inputs (Hendrycks et al. 2020). This is particularly important as $h_{\text{distinct}}^{\text{EHR}}$ and $h_{\text{distinct}}^{\text{CXR}}$ are generated from encoders with very different architectures from heterogeneous input, resulting in a highly diverse range of values. Formally, the loss function of shared representation alignment is given by

$$\mathcal{L}_{\text{JSD}} = \frac{1}{2}\left(\text{KL}(P||M) + \text{KL}(Q||M)\right), \tag{1}$$

where $M = (P + Q)/2$ denotes the mixture of P and Q, and KL denotes the Kullback–Leibler divergence. The logits corresponding to M then can be computed by $\sigma^{-1}(M)$, where $\sigma^{-1}(\cdot)$ denotes the logit function, the inverse of the standard logistic function. We define the process of obtaining the logits of the mixture of the induced distributions from $h_{\text{distinct}}^{\text{EHR}}$ and $h_{\text{distinct}}^{\text{CXR}}$ as *logit pooling*, given by:

Definition 1 (Logit Pooling)**.** The logit pooling of h_1 and h_2 is given by:

$$\begin{aligned} \text{LogitPool}(h_1, h_2) &= \sigma^{-1}\left(\frac{\sigma(h_1) + \sigma(h_2)}{2}\right) \\ &= \log\frac{2e^{h_1+h_2} + e^{h_1} + e^{h_2}}{2 + e^{h_1} + e^{h_2}}. \end{aligned} \tag{2}$$

Since the shared representations are aligned, when both modalities are present, we can obtain the final shared representation via logit pooling. On the other hand, when CXR is missing, we can directly use the shared representation extracted from EHR data as the final shared representation. That is,

$$h_{\text{shared}} = \begin{cases} \text{LogitPool}(h_{\text{shared}}^{\text{EHR}}, h_{\text{shared}}^{\text{CXR}}) & \text{if } \mathbf{X}^{\text{CXR}} \neq \emptyset, \\ h_{\text{shared}}^{\text{EHR}} & \text{otherwise.} \end{cases} \tag{3}$$

Representation disentanglement via orthogonality. The information shared across modalities and that is distinct within each modality are not naturally separated. To enable the modal-distinct ones to capture information that is not shared by the other modality, we impose orthogonality constraints to disentangle the modal-distinct information and reduce the redundancy in the shared and the modal-distinct representations (Jia et al. 2020). The orthogonality constraint can be enforced by minimizing the absolute value of the cosine similarities between the distinct representation and the shared representation for each modality. Formally, we have:

$$\begin{aligned} \mathcal{L}_{\text{orth}}^{\text{EHR}} &= \ell_{\text{orth}}(h_{\text{shared}}^{\text{EHR}}, h_{\text{distinct}}^{\text{EHR}}) \\ \mathcal{L}_{\text{orth}}^{\text{EHR}} &= \ell_{\text{orth}}(h_{\text{shared}}^{\text{CXR}}, h_{\text{distinct}}^{\text{CXR}}), \end{aligned} \tag{4}$$

where $\ell_{\text{orth}}(\mathbf{h}_1, \mathbf{h}_2) = \frac{|\langle \mathbf{h}_1, \mathbf{h}_2 \rangle|}{||\mathbf{h}_1||_2 \cdot ||\mathbf{h}_2||_2}$, and $\langle \mathbf{h}_1 \cdot \mathbf{h}_2 \rangle$ denotes the inner product between vectors $\mathbf{h}_1$ and $\mathbf{h}_2$.

Disease-aware Masked Attention Fusion

Inspired by the fact that clinicians rely on different diagnostic tools on varying scales according to the patient's health condition and the particular disease, we propose to learn the significance of each modal regarding predicting different diseases for different patients. To this end, we develop a disease-aware masked attention fusion module that could respect the importance of each modality for different prediction targets.

First, we compute the query vector by taking the average of the available representations followed by a linear projection, given by:

$$\mathbf{q} = \begin{cases} (\mathbf{h}_{\text{distinct}}^{\text{EHR}} + \mathbf{h}_{\text{shared}} + \mathbf{h}_{\text{distinct}}^{\text{CXR}})\mathbf{W}^Q/3 & \text{if } \mathbf{X}^{\text{CXR}} \neq \emptyset, \\ (\mathbf{h}_{\text{distinct}}^{\text{EHR}} + \mathbf{h}_{\text{shared}})\mathbf{W}^Q/2 & \text{otherwise.} \end{cases} \tag{5}$$

The query vector can be regarded as a summary of the medical status of the patient. To allow different modal significance to be captured, we compute a set of "target vectors", each corresponding to a particular prediction target:

$$\mathbf{K}_c = \mathbf{H}\mathbf{W}_c^K, \tag{6}$$

where c denotes the index of the prediction target and $\mathbf{H}$ is obtained by stacking the representations row-wisely:

$$\mathbf{H} = \left[(\mathbf{h}_{\text{distinct}}^{\text{EHR}})^\top, (\mathbf{h}_{\text{shared}})^\top, (\mathbf{h}_{\text{distinct}}^{\text{CXR}})^\top\right] \in \mathbb{R}^{3 \times d}.$$

We follow the scaled-product attention (Vaswani et al. 2017) to generate the attention weightings of the three representations for each prediction target:

$$\boldsymbol{\alpha}_c = \text{softmax}\left(\frac{\mathbf{q}\mathbf{K}_c + \mathbf{m}}{\sqrt{d}}\right), \quad c = 1, \ldots, |C|, \tag{7}$$

where $|C|$ is the number of prediction classes and $\mathbf{m} \in \{1, -\infty\}^3$ is a masking vector. It takes the value of ones except the third entry, m_3, which equals negative infinity when CXR is missing, one otherwise. The final representation for the cth prediction target is given by:

$$\tilde{\mathbf{h}}_c = \boldsymbol{\alpha}_c^\top \mathbf{H}\mathbf{W}^V, \tag{8}$$

where $\mathbf{W}^Q$, $\mathbf{W}_c^K$, and $\mathbf{W}^V$ are projection matrices.

Attention ranking loss. To further enforce the modal significance to be explicitly captured, we propose an attention ranking loss. First, we train auxiliary classifiers using $\mathbf{h}_{\text{distinct}}^{\text{EHR}}$, $\mathbf{h}_{\text{shared}}$, and $\mathbf{h}_{\text{distinct}}^{\text{CXR}}$ as input jointly with the model learning, producing three predictions:

$$\hat{\mathbf{y}}_1 = g_1(\mathbf{h}_{\text{distinct}}^{\text{EHR}}), \ \hat{\mathbf{y}}_2 = g_2(\mathbf{h}_{\text{shared}}), \ \text{and} \ \hat{\mathbf{y}}_3 = g_3(\mathbf{h}_{\text{distinct}}^{\text{CXR}}),$$

where g's are parameterized by two-layer feedforward networks. We use cross-entropy as the auxiliary loss function:

$$\mathcal{L}_{\text{aux}} = \sum_{i=1}^{3}\sum_{c=1}^{|C|}\ell_{ci} \tag{9}$$

with $\ell_{ci} = y_c \log(\hat{y}_{ci}) + (1 - y_c)\log(1 - \hat{y}_{ci})$.

The auxiliary loss function reflects the capability of each representation to predict the target. Thus, we enforce the attention weights to have a ranking consistent with the order of three loss values. We use the following margin ranking loss:

$$\mathcal{L}_{\text{attn}} = \frac{1}{2|C|}\sum_{c=1}^{|C|}\sum_{i=1}^{3}\sum_{j\neq i}\max\left(0, \mathbb{1}[\ell_{ci} < \ell_{cj}](\alpha_{cj} - \alpha_{ci}) + \epsilon\right), \tag{10}$$

where $\mathbb{1}[\cdot]$ is the indicator function. It equals one if the condition holds, zero otherwise. Eq. (10) imposes penalty when the prediction y_{ci} is better than y_{cj} but the attention weighting α_{ci} is not greater than α_{cj} with a margin of ϵ.
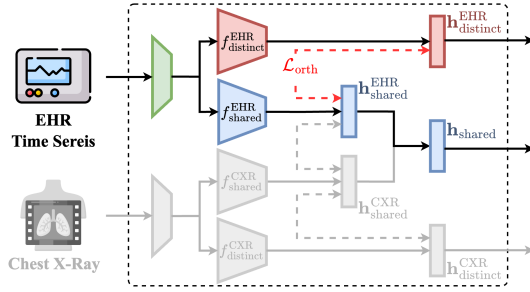
Figure 2: Data flow in the disentangled representation learning module when the CXR modality is missing. The shared representation extracted from EHR will be directly used as $\mathbf{h}_{\text{shared}}$. Inactive components and loss terms are grayed out.

Learning Algorithms

After obtaining the final representations $\tilde{\mathbf{h}}_c$, the final prediction for the cth class can be obtained using a feedforward layer: $\hat{y}_c = \psi_c(\tilde{\mathbf{h}}_c)$. The loss function for the final prediction can be given by cross entropy as:

$$\mathcal{L}_{\text{pred}} = \sum_{c=1}^{|C|} y_c \log(\hat{y}_c) + (1 - y_c)\log(1 - \hat{y}_c). \quad (11)$$

The overall loss function to minimize is then given by adding the distribution alignment loss in Eq. (1), the disentanglement loss in Eq. (4), the auxiliary loss in Eq. (9), the attention ranking loss in Eq. (10), and the final prediction loss in Eq. (11):

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_1 \mathcal{L}_{\text{JSD}} + \lambda_2(\mathcal{L}_{\text{orth}}^{\text{EHR}} + \mathcal{L}_{\text{orth}}^{\text{CXR}}) + \lambda_3(\mathcal{L}_{\text{attn}} + \mathcal{L}_{\text{aux}}). \quad (12)$$

Training with missing modality. When CXR is not available, we extract the disentangled representation from EHR data only and use the EHR shared representation as $\mathbf{h}_{\text{shared}}$ directly, as in Eq. (3), and loss terms in Eq. (12) involving CXR representations are removed. Therefore, the objective function to be optimized over the entire training set with partially missing CXR data is given by:

$$\min \quad \frac{1}{|\mathcal{D}|}\left(\sum_{i \in \mathcal{D}_{\text{paired}}} \mathcal{L}_i + \sum_{i \in \mathcal{D}_{\text{partial}}} \left(\mathcal{L}_{\text{pred}} + \lambda_2 \mathcal{L}_{\text{orth}}^{\text{EHR}}\right) \right), \quad (13)$$

where i is the index of patients.

Experiments

Experiment Settings

Datasets and preprocessing. We use the large-scale real-world EHR datasets, MIMIC-IV (Johnson et al. 2023) and MIMIC-CXR (Johnson et al. 2019) to empirically evaluate the predictive performance of DrFuse. MIMIC-IV contains de-identified data of adult patients admitted to either intensive care units or the emergency department of Beth Israel Deaconess Medical Center (BIDMC) between 2008 and 2019. MIMIC-CXR is a publicly available dataset of chest

	Missing Modality	Training	Validating	Testing
full dataset	✓	42,628	4,802	11,914
matched subset	✗	7,637	857	2,136

Table 1: Number of samples in the two datasets constructed.

radiographs collected from BIDMC, where a subset of patients can be matched with those in MIMIC-IV.

We follow similar procedures to preprocess the data as those in (Hayat, Geras, and Shamout 2022). We extract 17 clinical variables that are routinely monitored in ICU, including five categorical variables and twelve continuous ones. We use disease prediction as the prediction task, where the 25 disease phenotype labels are generated based on diagnosis codes following (Harutyunyan et al. 2019). To better align with the clinical need for early prediction, we make predictions of the disease phenotypes using data within the first 48 hours of the ICU admission. Accordingly, we retrieve the last Anterior-Posterior (PA) projection chest X-ray in the same observation window. In total, we extracted $59,344$ ICU stays with EHR records, of which $10,630$ are associated with CXR.

To test DrFuse in different modality missing settings, we construct two datasets using the extracted data: a *full dataset* containing all patients regardless of having CXR or not, and a *matched subset* only containing patients having both EHR and CXR. We randomly split the dataset with a ratio of 7:1:2 for training, validation, and testing. It is worth noting that patients in the validation and test subsets of the *matched subset* are also split into validation and test subsets, respectively. This allowed us to train the model using one dataset and test it with the other. Table 1 shows the number of patients having each data modality.

Evaluation metrics. Due to the highly imbalanced nature of the disease labels (see Table 3 for prevalence), we evaluate the performance of DrFuse and baseline models using Area Under the Precision-Recall Curve (PRAUC).

Experiment implementation. The experiment environment is a machine equipped with dual Intel Xeon Silver 4114 CPUs and four Nvidia V100 GPU cards. The model is implemented using Pytorch 2.0.1. We use grid search to tune the hyperparameters using the validation set and report that over the test set. The search spaces of the hyperparameters are: $\lambda_1 \in \{0, 0.1, \mathbf{1}\}$, $\lambda_2 \in \{0, 0.1, \mathbf{1}\}$, $\lambda_3 \in \{0, \mathbf{0.5}, 1\}$, lr $\in \{\mathbf{0.0001}, 0.001\}$, where the value in bold indicates the optimal choice. When training with the matched subset, we randomly remove the CXR of 30% samples within each mini-batch as an additional data augmentation.

Baseline models. We compare against the following baselines. **MMTM** (Joze et al. 2020) is a module that can leverage the information between modalities with flexible plug-in architectures. Since the model assumes full modality, we compensate for the missing modality CXR with all zeros during training and testing. **DAFT** (Pölsterl, Wolf, and Wachinger 2021) is a module that can be plugged into CNN

Model	Trained with the *matched subset*		Trained with the *full dataset*	
	testing on *matched subset*	testing on *full dataset*	testing on *matched subset*	testing on *full dataset*
Transformer	0.408 (0.368, 0.455)	0.374 (0.355, 0.395)	0.435 (0.393, 0.481)	0.418 (0.398, 0.440)
MMTM	0.416 (0.378, 0.462)	0.359 (0.342, 0.379)	0.422 (0.383, 0.469)	0.407 (0.387, 0.428)
DAFT	0.417 (0.376, 0.462)	0.348 (0.331, 0.368)	0.430 (0.389, 0.477)	0.409 (0.389, 0.431)
MedFuse	0.427 (0.387, 0.473)	0.329 (0.312, 0.347)	0.434 (0.394, 0.481)	0.405 (0.385, 0.427)
MedFuse-II	0.418 (0.378, 0.463)	0.329 (0.314, 0.348)	0.427 (0.387, 0.473)	0.412 (0.391, 0.433)
DrFuse	**0.450 (0.426, 0.498)**	**0.384 (0.371, 0.402)**	**0.470 (0.420, 0.512)**	**0.419 (0.391, 0.434)**

Table 2: Overall performance measured by the macro average of PRAUC over all 25 disease phenotype labels for different combinations of training and test subsets. Numbers in bold indicate the best performance in each column. `DrFuse` consistently outperforms all baselines in all settings with a significant margin.
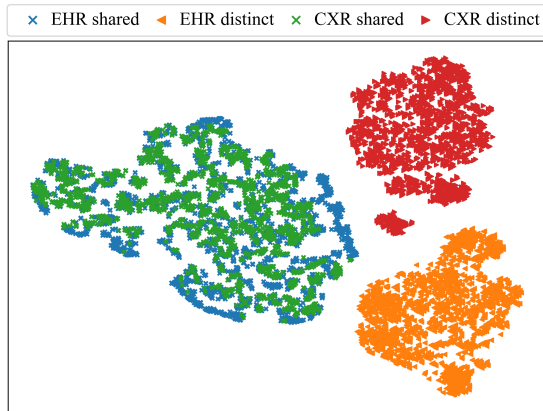

Figure 3: t-SNE visualization of distinct and shared features for the test set in the *matched subset*. `DrFuse` could well align the distributions of the EHR and CXR shared representations, as well as disentangle the distinct representations.

models to exchange information between tabular data and image modality. Similarly, we replace the input of CXR with matrices of all zeros during training and testing. **MedFuse** (Hayat, Geras, and Shamout 2022) uses an LSTM-based fusion to combine features from the image encoder and EHR encoder. Missing modality is handled by learning a global representation for the missing CXR. **MedFuse-II** is a variant of MedFuse with its CXR encoders and EHR encoders replaced by ResNet50 and Transformer, respectively, to ensure a fair comparison with `DrFuse`. **Transformer** (Vaswani et al. 2017) is the EHR encoder used by `DrFuse`, which is a uni-modal method that takes only EHR as input.

Overall Performance of Disease Prediction

The performance in terms of disease phenotype prediction is summarized in Table 2. We report the macro average of PRAUC over all 25 disease phenotype labels together with the corresponding 95% confidence interval obtained through 1000 iterations using the bootstrap method. The results show that `DrFuse` consistently outperforms all baselines compared with a large margin. When trained and tested both with the *matched subset*, *i.e.*, no missing modality

is involved, `DrFuse` achieves 5.4% relative improvement against MedFuse, demonstrating that the proposed `DrFuse` could achieve effective modality fusion. When trained with the *full dataset* and tested with the *matched subset*, `DrFuse` achieves 8% relative improvement against MedFuse, suggesting that `DrFuse` could fully utilize the training samples with missing modalities. When tested on the *full dataset*, all methods, including the uni-modal Transformer, obtain worse results compared with the test scores obtained on the *matched subset*. This suggests that a severe domain shift could exist between the two subsets. This might be because patients who could not undergo X-ray scans may have more complex health conditions and thus are much harder to predict. Having said so, `DrFuse` still obtains the best performance in the presence of such potential severe domain shift in the *full dataset* benefited from the representation disentanglement.

Disease-Wise Prediction Performance

To gain more insights into the prediction performance, we show the disease-wise PRAUC scores obtained by the uni-modal methods, MedFuse, and `DrFuse` in Table 3. Numbers inside parentheses indicate the relative difference against the best uni-modal prediction. The results show that combining EHR and CXR is not always helpful for all diseases, due to the modal inconsistency issue as mentioned earlier. For example, when predicting *conduction disorders* and *other upper respiratory disease*, the performance of MedFuse drops 40.5% and 30.9%, respectively, compared with uni-modal predictions. On the contrary, `DrFuse` only drops 1% for *conduction disorders* and achieves 24.2% improvement for *other upper respiratory disease*. This demonstrates that the proposed `DrFuse` could better address the modal inconsistency issue by inferring the disease-specific and patient-specific modal significance.

Visualization of Disentangled Representation

To further validate the effectiveness of the disentangled representation learning, we visualize the shared and distinct representations for EHR and CXR data with t-SNE in Fig. 3. The shared representations, $\mathbf{h}_{\text{shared}}^{\text{EHR}}$ and $\mathbf{h}_{\text{shared}}^{\text{CXR}}$, are well blended as a cluster. Meanwhile the distinct representations, $\mathbf{h}_{\text{distinct}}^{\text{EHR}}$ and $\mathbf{h}_{\text{distinct}}^{\text{CXR}}$, remain well-separated not just from each other but also from the shared features.

Disease Label	Prevalence	ResNet50 (CXR)	Transformer (EHR)	MedFuse	MedFuse-II	DrFuse
Acute and unspecified renal failure	0.32	0.469	0.537	0.559 (4.1%)	0.541 (0.7%)	0.541 (0.7%)
Acute cerebrovascular disease	0.07	0.145	0.457	0.461 (0.9%)	0.441 (-3.5%)	0.441 (-3.5%)
Acute myocardial infarction	0.09	0.165	0.170	**0.217 (27.6%)**	0.177 (4.1%)	**0.193 (13.5%)**
Cardiac dysrhythmias	0.38	0.566	0.513	0.552 (-2.5%)	0.517 (-8.7%)	0.568 (0.4%)
Chronic kidney disease	0.24	0.400	0.424	**0.455 (7.3%)**	**0.455 (7.3%)**	**0.445 (5%)**
Chronic obstructive pulmonary disease	0.15	0.374	0.239	0.323 (-13.6%)	0.317 (-15.2%)	0.355 (-5.1%)
Complications of surgical/medical care	0.22	0.303	0.408	0.379 (-7.1%)	0.395 (-3.2%)	0.407 (-0.2%)
Conduction disorders	0.11	0.625	0.237	0.372 (-40.5%)	0.231 (-63%)	0.619 (-1%)
Congestive heart failure; nonhypertensive	0.29	0.593	0.509	0.597 (0.7%)	0.558 (-5.9%)	**0.629 (6.1%)**
Coronary atherosclerosis and related	0.34	0.657	0.559	0.603 (-8.2%)	0.588 (-10.5%)	0.640 (-2.6%)
Diabetes mellitus with complications	0.12	0.217	0.520	0.469 (-9.8%)	0.505 (-2.9%)	0.486 (-6.5%)
Diabetes mellitus without complication	0.21	0.276	0.361	0.338 (-6.4%)	0.363 (0.6%)	**0.381 (5.5%)**
Disorders of lipid metabolism	0.41	0.587	0.593	0.598 (0.8%)	0.612 (3.2%)	0.612 (3.2%)
Essential hypertension	0.44	0.558	0.578	0.592 (2.4%)	0.601 (4%)	0.572 (-1%)
Fluid and electrolyte disorders	0.45	0.563	0.675	0.675 (0%)	0.663 (-1.8%)	0.660 (-2.2%)
Gastrointestinal hemorrhage	0.07	0.121	0.193	0.152 (-21.2%)	0.152 (-21.2%)	**0.204 (5.7%)**
Hypertension with complications	0.22	0.378	0.393	**0.418 (6.4%)**	**0.424 (7.9%)**	0.409 (4.1%)
Other liver diseases	0.17	0.341	0.268	0.351 (2.9%)	0.319 (-6.5%)	**0.389 (14.1%)**
Other lower respiratory disease	0.13	0.182	0.170	0.167 (-8.2%)	0.176 (-3.3%)	0.186 (2.2%)
Other upper respiratory disease	0.05	0.102	0.165	0.114 (-30.9%)	0.161 (-2.4%)	**0.205 (24.2%)**
Pleurisy; pneumothorax; pulmonary collapse	0.10	0.195	0.126	0.191 (-2.1%)	0.156 (-20%)	0.192 (-1.5%)
Pneumonia	0.18	0.354	0.404	0.400 (-1%)	**0.428 (5.9%)**	0.419 (3.7%)
Respiratory failure; insufficiency; arrest (adult)	0.28	0.520	0.607	0.591 (-2.6%)	0.605 (-0.3%)	0.615 (1.3%)
Septicemia (except in labor)	0.22	0.371	0.538	0.522 (-3%)	0.514 (-4.5%)	0.528 (-1.9%)
Shock	0.17	0.342	0.558	0.567 (1.6%)	0.545 (-2.3%)	0.542 (-2.9%)
Average Rank		3.96	3.24	2.68	2.84	**2.04**

Table 3: The PRAUC score for each disease label. "ResNet50" and "Transformer" indicate the performance obtained using only CXR data and EHR data, respectively. The percentages within parentheses indicate the relative difference against the best uni-modal prediction. Results show that DrFuse could better address the inconsistency issue, resulting in the highest average rank over all disease labels. Results with relative differences beyond ±5% over the best uni-modal predictions are highlighted.

Model	PRAUC @matched subset	PRAUC @full dataset
w/o disentangled	0.446 (0.411, 0.501)	0.374 (0.355, 0.395)
MSE alignment	0.447 (0.410, 0.498)	0.375 (0.356, 0.396)
w/o attn. ranking	0.438 (0.396, 0.485)	0.361 (0.343, 0.382)
DrFuse	0.450 (0.426, 0.498)	0.384 (0.371, 0.402)

Table 4: Results of the ablation study tested over different datasets by removing each component from DrFuse. The models are trained using the *matched subset*.

Ablation Study

To gain further insights into the source of performance gain of DrFuse, we conduct an ablation study by training the model using the *matched subset* with each component of DrFuse being removed. The results are summarized in Table 4. The first row is obtained by removing $\mathcal{L}_{JSD}$ and $\mathcal{L}_{orth}$ and the second row is obtained by replacing the JSD with the MSE loss and the logit pooling with the average pooling. A significant performance drop can be observed when the model is tested using the *full dataset*. This shows that the proposed disentangled representation learning is effective in handling missing modalities. The third row removes the at-

tention ranking loss $\mathcal{L}_{attn}$, where a significant performance drop is observed, showing that capturing disease-wise modal significance is important for the disease prediction task and the proposed method is effective in achieving this goal.

Conclusion

In this paper, we propose a novel model, DrFuse, that learns the disentangled representation from EHR and CXR data to achieve medical multi-modal data fusion in the presence of missing modality and modal inconsistency. A shared representation and a distinct representation are learned from each modal. We align the shared representations by minimizing the Jensen–Shannon divergence (JSD) and achieve representation disentanglement via imposing orthogonal constraints. A logit pooling operation is derived to fuse the shared representations. Besides, we propose a disease-aware attention fusion module that captures the patient-specific modal significance for each prediction target via an attention ranking loss. The experimental results demonstrate that the proposed model is effective in achieving disentangled representation, addressing the missing modality and modal inconsistency issues. For future research, we will focus on the domain shift between patients with and without CXR jointly with multi-modal learning.

Acknowledgements

References

Ahmad, J. 2016. The diabetic foot. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 10(1): 48–60.

Aljondi, R.; and Alghamdi, S. 2020. Diagnostic value of imaging modalities for COVID-19: scoping review. *Journal of Medical Internet Research*, 22(8): e19673.

Brouwer, M. C.; Tunkel, A. R.; and van de Beek, D. 2010. Epidemiology, diagnosis, and antimicrobial treatment of acute bacterial meningitis. *Clinical Microbiology Reviews*, 23(3): 467–492.

Chen, C.; Dou, Q.; Jin, Y.; Chen, H.; Qin, J.; and Heng, P.-A. 2019. Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*, 447–456. Springer.

Harutyunyan, H.; Khachatrian, H.; Kale, D. C.; Ver Steeg, G.; and Galstyan, A. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1): 96.

Hayat, N.; Geras, K. J.; and Shamout, F. E. 2022. MedFuse: Multi-modal fusion with clinical time-series data and chest X-ray images. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, 479–503. PMLR.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hendrycks, D.; Mu, N.; Cubuk, E. D.; Zoph, B.; Gilmer, J.; and Lakshminarayanan, B. 2020. AugMix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*.

Hoare, Z.; and Lim, W. S. 2006. Pneumonia: update on diagnosis and management. *BMJ*, 332(7549): 1077–1079.

Huang, S.-C.; Pareek, A.; Seyyedi, S.; Banerjee, I.; and Lungren, M. P. 2020a. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ Digital Medicine*, 3(1): 136.

Huang, S.-C.; Pareek, A.; Zamanian, R.; Banerjee, I.; and Lungren, M. P. 2020b. Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. *Scientific Reports*, 10(1): 22147.

Jia, X.; Jing, X.-Y.; Zhu, X.; Chen, S.; Du, B.; Cai, Z.; He, Z.; and Yue, D. 2020. Semi-supervised multi-view deep discriminant representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7): 2496–2509.

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

Johnson, A. E.; Pollard, T. J.; Berkowitz, S. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Mark, R. G.; and Horng, S. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1): 317.

Joze, H. R. V.; Shaban, A.; Iuzzolino, M. L.; and Koishida, K. 2020. MMTM: Multimodal transfer module for CNN fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13289–13299.

Kline, A.; Wang, H.; Li, Y.; Dennis, S.; Hutch, M.; Xu, Z.; Wang, F.; Cheng, F.; and Luo, Y. 2022. Multimodal machine learning in precision health: A scoping review. *NPJ Digital Medicine*, 5(1): 171.

Li, L.; Ding, W.; Huang, L.; Zhuang, X.; and Grau, V. 2023. Multi-modality cardiac image computing: A survey. *Medical Image Analysis*, 102869.

Lin, M.; Wang, S.; Ding, Y.; Zhao, L.; Wang, F.; and Peng, Y. 2021. An empirical study of using radiology reports and images to improve ICU-mortality prediction. In *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, 497–498. IEEE.

Ma, M.; Ren, J.; Zhao, L.; Tulyakov, S.; Wu, C.; and Peng, X. 2021. SMIL: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2302–2310.

Mohsen, F.; Ali, H.; El Hajj, N.; and Shah, Z. 2022. Artificial intelligence-based methods for fusion of electronic health records and imaging data. *Scientific Reports*, 12(1): 17981.

Pölsterl, S.; Wolf, T. N.; and Wachinger, C. 2021. Combining 3D image and tabular data via the dynamic affine feature map transform. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, 688–698. Springer.

Sharma, A.; and Hamarneh, G. 2019. Missing MRI pulse sequence synthesis using multi-modal generative adversarial network. *IEEE Transactions on Medical Imaging*, 39(4): 1170–1183.

Shen, Y.; and Gao, M. 2019. Brain tumor segmentation on MRI with missing modalities. In *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*, 417–428. Springer.

Sun, M.; Zhang, X.; Ma, J.; Xie, S.; Liu, Y.; and Philip, S. Y. 2023. Inconsistent matters: A knowledge-guided dual-consistency network for multi-modal rumor detection. *IEEE Transactions on Knowledge and Data Engineering*.

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

Venugopalan, J.; Tong, L.; Hassanzadeh, H. R.; and Wang, M. D. 2021. Multimodal deep learning models for early detection of Alzheimer's disease stage. *Scientific reports*, 11(1): 3254.

Wang, H.; Chen, Y.; Ma, C.; Avery, J.; Hull, L.; and Carneiro, G. 2023. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15878–15887.

Xiong, S.; Zhang, G.; Batra, V.; Xi, L.; Shi, L.; and Liu, L. 2023. TRIMOON: Two-round inconsistency-based multimodal fusion network for fake news detection. *Information Fusion*, 93: 150–158.

Yoo, Y.; Tang, L. Y.; Li, D. K.; Metz, L.; Kolind, S.; Traboulsee, A. L.; and Tam, R. C. 2019. Deep learning of brain lesion patterns and user-defined clinical and MRI features for predicting conversion to multiple sclerosis from clinically isolated syndrome. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 7(3): 250–259.

Zarogoulidis, P.; Kioumis, I.; Pitsiou, G.; Porpodis, K.; Lampaki, S.; Papaiwannou, A.; Katsikogiannis, N.; Zaric, B.; Branislav, P.; Secen, N.; et al. 2014. Pneumothorax: from definition to diagnosis and treatment. *Journal of Thoracic Disease*, 6(Suppl 4): S372.

Zhang, C.; Chu, X.; Ma, L.; Zhu, Y.; Wang, Y.; Wang, J.; and Zhao, J. 2022. M3Care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2418–2428.

Zhao, Z.; Yang, H.; and Sun, J. 2022. Modality-adaptive feature interaction for brain tumor segmentation with missing modalities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 183–192. Springer.

Zhi, Z.; Elbadawi, M.; Daneshmend, A.; Orlu, M.; Basit, A.; Demosthenous, A.; and Rodrigues, M. 2022. Multimodal diagnosis for pulmonary embolism from EHR data and CT images. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2053–2057. IEEE.