MedFuse: Multi-modal fusion with clinical time-series data and chest X-ray images

Nasir Hayat   nasirhayat6160@gmail.com
G42 Healthcare
Abu Dhabi, UAE

Krzysztof J. Geras   k.j.geras@nyu.edu
Department of Radiology
NYU Grossman School of Medicine
New York, NY, USA

Farah E. Shamout   farah.shamout@nyu.edu
Engineering Division
NYU Abu Dhabi
Abu Dhabi, UAE

Abstract

Multi-modal fusion approaches aim to integrate information from different data sources. Unlike natural datasets, such as in audio-visual applications, where samples consist of "paired" modalities, data in healthcare is often collected asynchronously. Hence, requiring the presence of all modalities for a given sample is not realistic for clinical tasks and significantly limits the size of the dataset during training. In this paper, we propose `MedFuse`, a conceptually simple yet promising LSTM-based fusion module that can accommodate uni-modal as well as multi-modal input. We evaluate the fusion method and introduce new benchmark results for in-hospital mortality prediction and phenotype classification, using clinical time-series data in the MIMIC-IV dataset and corresponding chest X-ray images in MIMIC-CXR. Compared to more complex multi-modal fusion strategies, `MedFuse` provides a performance improvement by a large margin on the fully paired test set. It also remains robust across the partially paired test set containing samples with missing chest X-ray images. We release our code for reproducibility and to enable the evaluation of competing models in the future.

1. Introduction

Humans perceive the world through multi-modal data (Ngiam et al., 2011). To date, most of the successful models learning from perceptual data in healthcare are uni-modal, i.e. they rely on a single data modality (Huang et al., 2020a). Multi-modal learning has been widely explored in the context of audio-visual applications (Vaezi Joze et al., 2020) and natural image datasets (Zellers et al., 2021; Hayat et al., 2020), but less so in healthcare. The main goal of multi-modal fusion is to exploit relevant information from different modalities to improve performance in downstream tasks (Baltrušaitis et al., 2018). Multi-modal fusion strategies can be characterized as early, joint, or late fusion (Huang et al., 2020a). The joint fusion paradigm is the most promising, since its core idea is to model interactions between the representations of the input modalities.

We highlight two main challenges facing multi-modal joint fusion in healthcare. First, many of the state-of-the-art approaches make a strong assumption that all modalities are available for every sample during training, inference, or both (Pölsterl et al., 2021). Although some clinical studies follow suit of this assumption (Huang et al., 2020a), obtaining paired data is not feasible since daily clinical practice produces heterogeneous data with varying sparsity. For example, physiological data is more frequently collected than chest X-ray images in the Intensive Care Unit (ICU) setting. These two modalities are the key focus of our study because they play a very important role in clinical prediction tasks (Harutyunyan et al., 2019; Lohan, 2019). Developing a unified fusion model for those two modalities also presents its own challenges as they (i) have significantly different input dimensions, (ii) require modality-specific feature extractors due to differences in information and noise content (Nagrani et al., 2021), and (iii) are not temporally aligned and hence cannot be paired easily. Considering those challenges, our primary aim is to propose a fusion architecture that can deal with partially paired data, in order to achieve favorable performance in downstream prediction tasks.

The second challenge is that there are no well-studied publicly available multi-modal clinical benchmarks. Therefore, most studies rely on a single data modality to perform clinical prediction tasks (Harutyunyan et al., 2019), or use privately curated multi-modal datasets (Huang et al., 2020a). Here, our secondary aim is to introduce new multi-modal benchmark results for two popular clinical prediction tasks using the publicly available Medical Information Mart for Intensive Care (MIMIC)-IV (Johnson et al., 2021) and MIMIC-CXR (Johnson et al., 2019) datasets, and we also release the code for reproducibility. We compare our approach to vanilla early and joint fusion as well as open-source state-of-the-art joint fusion approaches (Vaezi Joze et al., 2020; Pölsterl et al., 2021). In summary, we make the following contributions:

- We propose `MedFuse`, a new LSTM-based (Hochreiter and Schmidhuber, 1997) multi-modal fusion approach. Conventional joint fusion strategies concatenate feature representations of multiple modalities as a single feature representation, and then process that concatenated representation for downstream tasks, such as using a classifier. On the contrary, we treat the multi-modal representation as a sequence of uni-modal representations (or tokens), such that the fusion module aggregates these representations through the recurrence mechanism of LSTM. We assume a sequential structure to leverage the recurrent inductive bias of LSTM and to handle input sequences of variable length, in case of a missing modality. The fusion module is agnostic to the architecture of the modality-specific extractors and can handle missing data during training and inference.

- To evaluate the proposed approach, we link two open-access real-world datasets: MIMIC-IV (Johnson et al., 2021), which contains clinical time-series data collected in the ICU, and MIMIC-CXR (Johnson et al., 2019), which contains chest X-ray images. We pre-process the data and introduce new benchmark results for two tasks (Harutyunyan et al., 2019): in-hospital mortality prediction and phenotype classification. The results show that the model's performance remains robust across uni-modal samples and improves for paired multi-modal samples. The model achieves state-of-the-art results without imposing any assumptions on correlation between modalities.

- Considering the lack of multi-modal learning benchmarks in healthcare, we release our data pre-processing and benchmark code to allow reproducibility of the results and enable the evaluation of competing models in the future. The code can be found at: https://github.com/nyuad-cai/MedFuse. An overview of the proposed work is shown in Figure 1.
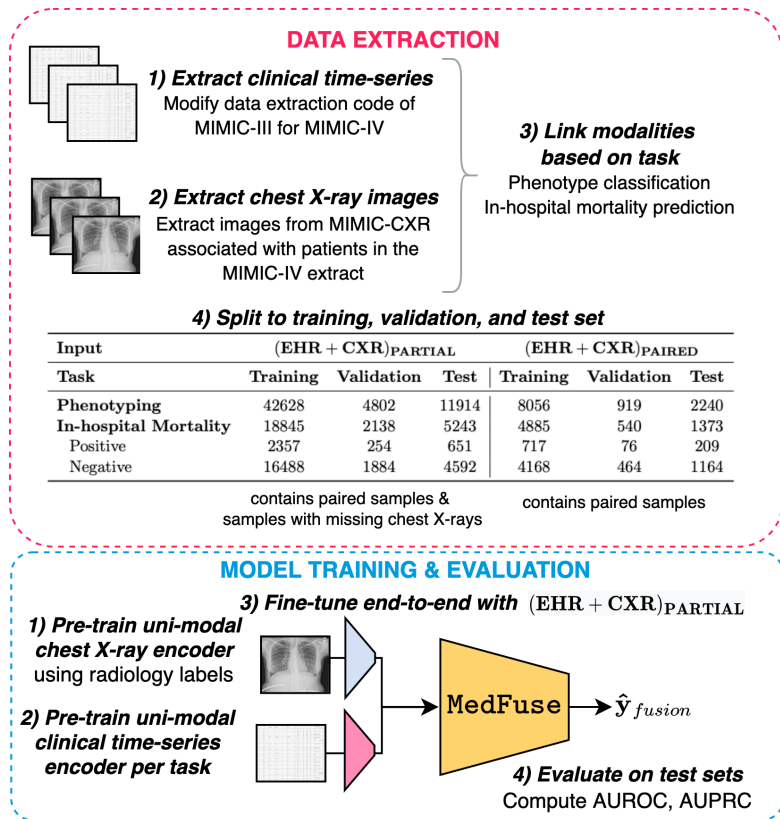

Figure 1: **Overview of the proposed work.** We first extract and link the datasets from MIMIC-IV and MIMIC-CXR based on the task definition (i.e., inhospital mortality prediction, or phenotype classification). The data splits of the training, validation, and test sets are summarized for each task, and the prevalence of positive and negative labels for in-hospital mortality is shown. Phenotype classification involves 25 labels as shown in Table 4.

Generalizable Insights about Machine Learning in the Context of Healthcare

State-of-the-art multi-modal fusion approaches typically investigate synchronous sources of information using natural datasets, such as audio, visual, and textual modalities. In healthcare, data is often sparse and heterogeneous and hence modalities are not always paired. Our work overcomes the challenge of missing data by proposing a flexible fusion approach that is agnostic to the modality-specific encoders. Therefore, it can be used for other types of input data, beyond chest X-ray images and clinical time-series data. It also highlights the value of processing a sequence of uni-modal representations, compared to the conventional concatenation strategy in joint fusion. Overall, the work highlights the promise of multi-modal fusion in healthcare to improve performance in downstream tasks.

2. Related Work

Routine clinical practice produces large amounts of data from different sources (i.e. modalities), including medical images, laboratory test results, measurements of vital signs, and clinical notes (Asri et al., 2015). Advances in deep learning have enabled building predictive models using subsets of modalities, typically clinical time-series data (Shickel et al., 2017) and medical images (Litjens et al., 2017). Here, we provide an overview of related work on multi-modal fusion in healthcare using imaging and non-imaging data.

2.1. Multi-modal Learning

Multi-modal learning has been widely explored for jointly learning representations of multiple modalities (Baltrušaitis et al., 2018). Example tasks include visual grounding (Chen et al., 2021), language grounding through visual cues (Zhang et al., 2021b), action recognition (Chen et al., 2015), video classification (Nagrani et al., 2021), image captioning (Yu et al., 2019), or visual-question answering (Zellers et al., 2021). Since machine learning studies typically investigate different combinations of audio, visual, and textual modalities, many of the existing methods are driven by the assumption that the modalities share intrinsic and structural information. This is not always true for heterogeneous data in healthcare. Hence, due consideration should be given to learning with multiple medical data modalities, since conventional assumptions for non-medical data are not necessarily applicable.

2.2. Multi-modal Fusion with Medical Images

There is an increasing interest in advancing the fusion of multi-modal medical images (Hermessi et al., 2021). The images usually represent different views of the same organ or lesion of interest, acquired using one or more sensors, whereby the images share the same set of labels. Proposed methods mainly focus on pixel-level fusion of complementary views acquired through multiple sensors to obtain a unified composite representation of the raw images (Li et al., 2021; James and Dasarathy, 2014). Various feature- and prediction-level fusion approaches were proposed for improved classification (Puyol-Antón et al., 2021; Wu et al., 2019; Zhang et al., 2021a) or segmentation performance (Hermessi et al., 2021). Since textual reports are a natural byproduct of radiology exams, they were also used as additional modalities for tasks like visual-question answering (Li et al., 2020; Sharma et al., 2021), report generation (Sonsbeek and Worring, 2020), or zero-shot image classification (Hayat et al., 2021b; Paul et al., 2021).

2.3. Multi-modal Fusion with Clinical Data and Medical Images

Several studies investigated the fusion of medical images and clinical data extracted from the patient's Electronic Health Records (EHR) for various applications (Huang et al., 2020a). For example, a stream of work covers tasks pertaining to cancer, such as recurrence prediction (Ho et al., 2021), lesion detection (Shao et al., 2020), or patient survival prediction (Vale-Silva and Rohr, 2020). Other tasks include detection of pulmonary embolism (Huang et al., 2020b), predicting the progression of Alzheimer's disease (Lee et al., 2019), diagnosis of neurological disease (Xin et al., 2021), or diagnosis of cervical dysplasia (Xu et al., 2016). While these studies highlight the impact of using multiple data
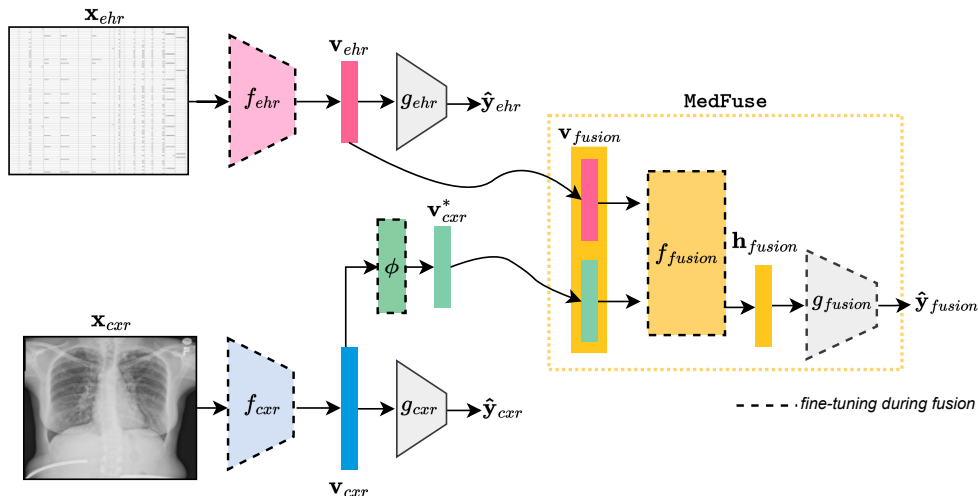
Figure 2: **Overview of network with MedFuse module.** First, we pre-train the modality-specific encoders and classifiers independently for each input modality. Specifically, we train f_{ehr} and g_{ehr} using the clinical time-series data and f_{cxr} and g_{cxr} using the chest X-ray images. Next, we project the chest X-ray latent representation $\mathbf{v}_{cxr}$ to $\mathbf{v}^*_{cxr}$, in order to match the dimension of $\mathbf{v}_{ehr}$. We pass $\mathbf{v}_{ehr}$ and $\mathbf{v}^*_{cxr}$ as an input sequence to the LSTM-based f_{fusion}, and we classify its last hidden state $\mathbf{h}_{fusion}$ to compute the overall prediction $\hat{\mathbf{y}}_{fusion}$. f_{fusion}, f_{ehr}, f_{cxr}, g_{fusion}, and ϕ are fine-tuned together for fusion.

modalities on downstream performance, many curate datasets for specific tasks and share the assumption that the images and selected clinical features are paired.

Some studies specifically focused on the integration of clinical data and chest X-ray images. For example, the integration of the two modalities showed a favorable impact on the predictive performance in prognostication tasks among patients with COVID-19 (Shamout et al., 2021; Jiao et al., 2021). Some studies jointly refine a common latent representation after aggregating encoded features of each modality (Grant et al., 2021; Jiao et al., 2021), while others combine predictions computed by each modality through weighted averaging (i.e., late fusion) (Shamout et al., 2021; Jiao et al., 2021). While late fusion enables the computation of predictions even for incomplete samples, it requires that the two modalities are assigned the same labels, which is not always feasible. Closely related to our work is that of Hayat et al. (2021a), where they propose a dynamic training approach for partially paired clinical time-series data and chest X-ray images for the task of phenotype classification. However, their method is not scalable since it incorporates an additional classifier (and prediction) for every possible combination of input modalities.

3. Methodology

We define a two stage-approach (i) to learn modality-specific perceptual models to extract the latent features (Section 3.1), and (ii) integrate these features through a joint multi-modal fusion module, MedFuse (Section 3.2). The overall architecture is shown in Figure 2. Without loss of generality, we focus here on two modalities only and denote the clinical time-series data as *ehr* and the chest X-ray images as *cxr* when defining the methodology.

3.1. Modality-specific Encoders

One of the main sources of heterogeneity in healthcare is the varying dimensionality of the input modalities, which makes it challenging to develop a unified encoder for all input modalities. Another difference is the target space, since we do not assume that the modalities must be assigned the same set of labels. Hence, we first define modality-specific encoders as follows.

For a given instance, let $\mathbf{x}_{ehr} \in \mathbb{R}^{t \times d}$ represent the clinical time-series data associated with ground-truth labels $\mathbf{y}_{ehr}$, where t is the number of time steps and d is the number of features derived from the clinical variables. We implement the encoder, f_{ehr}, for the clinical time-series modality as two stacked layers of an LSTM network (Hochreiter and Schmidhuber, 1997) with a dropout layer. We compute a latent feature representation $\mathbf{v}_{ehr} \in \mathbb{R}^m$ consisting of the last hidden state of the stacked LSTM, where $m = 256$. We then apply a classifier, g_{ehr}, to compute the predictions, such that $\hat{\mathbf{y}}_{ehr} = g_{ehr}(\mathbf{v}_{ehr})$. To fine-tune the encoder, we optimize the following loss:

$$\mathbb{L}_{ehr}(\mathbf{y}_{ehr}, \hat{\mathbf{y}}_{ehr}) = BCE(\mathbf{y}_{ehr}, \hat{\mathbf{y}}_{ehr}), \tag{1}$$

where BCE is the Binary Cross-Entropy loss.

Let $\mathbf{x}_{cxr} \in \mathbb{R}^{w \times h \times c}$ represent the chest X-ray image belonging to the same instance associated with the ground-truth labels $\mathbf{y}_{cxr}$, where w is the width dimension, h is the height dimension, and c is the number of channels. In all of our experiments, $h = 224$, $w = 224$, and $c = 3$, as we replicate each image across three channels. We implement the encoder, f_{cxr}, as a ResNet-34 (He et al., 2016) to compute $\mathbf{v}_{cxr} \in \mathbb{R}^n$, which is the feature representation after the average pooling layer of the convolutional network where $n = 512$. Similarly, we then apply a classifier, g_{cxr}, to compute the predictions, such that $\hat{\mathbf{y}}_{cxr} = g_{cxr}(\mathbf{v}_{cxr})$ and optimize the following loss to fine-tune the encoder:

$$\mathbb{L}_{cxr}(\mathbf{y}_{cxr}, \hat{\mathbf{y}}_{cxr}) = BCE(\mathbf{y}_{cxr}, \hat{\mathbf{y}}_{cxr}). \tag{2}$$

The encoders can hence be independently pre-trained using their respective labels and losses.

3.2. The `MedFuse` Module

To fuse the modalities, we first dismiss the classifiers, g_{ehr} and g_{cxr}, and keep the the pre-trained modality-specific encoders, f_{ehr} and f_{cxr}. Since the latent space dimensions of the two modalities are different, we use a projection layer, ϕ, that projects $\mathbf{v}_{cxr}$ to the same dimensionality as $\mathbf{v}_{ehr}$:

$$\mathbf{v}_{cxr}^* = \phi(\mathbf{v}_{cxr}) \tag{3}$$

such that $\mathbf{v}_{\mathbf{cxr}}^* \in \mathbb{R}^m$. We then create an input sequence consisting of the the uni-modal feature representations of the sample:

$$\mathbf{v}_{fusion} = [\mathbf{v}_{ehr}, \mathbf{v}_{cxr}^*]. \tag{4}$$

We parameterize a multi-modal fusion network, f_{fusion}, as a single LSTM layer with input dimension of 256 and a hidden dimension of 512, that aggregates the multi-modal sequence through recurrence. The motivation for using an LSTM is two-fold. First, it follows the

intuition of decision-making, where clinicians examine information from each modality sequentially, or one at a time. This allows the LSTM module to initially learn from $\mathbf{v}_{ehr}$, and then update its internal state using information in $\mathbf{v}_{cxr}^*$. Second, it can handle input sequences of variable number of modalities, so it inherently deals with missing modalities. In the case that the chest X-ray image is missing during training or inference, the network processes a single-element sequence, $[\mathbf{v}_{ehr}]$.

The last hidden state, $\mathbf{h}_{fusion}$, of f_{fusion} is then processed using a classifier g_{fusion} that computes the final fusion predictions, such that $\hat{\mathbf{y}}_{fusion} = g_{fusion}(\mathbf{h}_{fusion})$. We jointly train the encoders f_{ehr} and f_{cxr}, the projection layer ϕ, the fusion module f_{fusion}, and the classifier g_{fusion}, by optimizing the following loss:

$$\mathbb{L}_{fusion}(\mathbf{y}_{fusion}, \hat{\mathbf{y}}_{fusion}) = BCE(\mathbf{y}_{fusion}, \hat{\mathbf{y}}_{fusion}), \tag{5}$$

where $\mathbf{y}_{fusion} = \mathbf{y}_{ehr}$, since we assume that the clinical time-series data modality is the base modality associated with the prediction task of interest, and is always present during training and inference. All classifiers g_{ehr}, g_{cxr}, and g_{fusion} consist of a single linear layer followed by sigmoid activation.

4. Experiments

4.1. Datasets and Benchmark Tasks

For our experiments, we extract the clinical time-series data from MIMIC-IV (Johnson et al., 2021) along with the associated chest X-ray images in MIMIC-CXR (Johnson et al., 2019). Here we describe the two tasks and provide more details on each:

- **Phenotype classification:** The goal of this multi-label classification task is to predict whether a set of 25 chronic, mixed, and acute care conditions are assigned to a patient in a given ICU stay. For a given instance, $\mathbf{x}_{ehr}$ contains clinical time-series data collected during the entire ICU record, and $\mathbf{y}_{ehr}$ is a vector of 25 binary phenotype labels. We link each instance with the last chest X-ray image collected during the same ICU stay. MIMIC-III contains International Classification of Diseases (ICD) version 9 (ICD-9) codes, whereas MIMIC-IV contains both ICD-9 and ICD-10. In the original benchmark paper (Harutyunyan et al., 2019), the 25 phenotype labels were initially defined using the Clinical Classifications Software (CCS) for ICD-9 (WHO et al., 1988). Since ICD-9 and ICD-10 codes are aggregated to different CCS categories, we mapped all ICD-10 codes to ICD-9 using the guidelines provided by the Centers for Medicare & Medicaid Services[1], and then map them to CCS categories. We evaluate this task using the Area Under the Receiver Operating Characteristic (AUROC) curve and the Area Under the Precision Recall curve (AUPRC).
- **In-hospital mortality prediction:** The goal of this binary classification task is to predict in-hospital mortality after the first 48 hours spent in the ICU. Hence, for a given instance, $\mathbf{x}_{ehr}$ contains clinical time-series data collected during the first 48 hours of the ICU record, and $\mathbf{y}_{ehr}$ is a binary label indicating in-hospital mortality. Since the task requires a minimum of 48 hours, we exclude ICU stays that are shorter than

---

1. Centers for Medicare & Medicaid Services, https://www.cms.gov/Medicare/Coding/ICD10/2018-ICD-10-CM-and-GEMs

48 hours. Here, we pair each instance with the last chest X-ray image collected during the first 48 hours of ICU stay. We evaluate this task using AUROC and AUPRC.

4.1.1. Pre-processing of Clinical Time-series Data

We modified the extraction and data pre-possessing pipeline of Harutyunyan et al. (2019), which was originally implemented in TensorFlow (Abadi et al., 2015), and introduce a new version for MIMIC-IV using Pytorch (Paszke et al., 2019). To make a fair comparison and illustrate the efficacy of multi-modal learning, we use the same set of 17 clinical variables. Amongst those, five are categorical (capillary refill rate, Glasgow coma scale eye opening, Glasgow coma scale motor response, Glasgow coma scale verbal response, and Glasgow coma scale total) and 12 are continuous (diastolic blood pressure, fraction of inspired oxygen, glucose, heart rate, height, mean blood pressure, oxygen saturation, respiratory rate, systolic blood pressure, temperature, weight, and pH). For all the tasks, we regularly sample the input every two hours, discretize and standardize the clinical variables to obtain the input for f_{ehr} as in previous work (Harutyunyan et al., 2019). After data pre-processing and one-hot encoding of the categorical features, we obtain a vector representation of size 76 at each time-step of the clinical time-series data, such that for a given instance, $\mathbf{x}_{ehr} \in \mathbb{R}^{t \times 76}$ and t depends on the instance and task.

4.1.2. Data Splits

Using the patient identifier of the clinical time-series data, we randomly split the dataset into 70% for training, 10% for validation, and 20% for test set, as shown in Figure 1. We report final results on the test sets and compute 95% confidence intervals with 1000 iterations via the bootstrap method (Efron and Tibshirani, 1994). Here, we denote the clinical time-series data as **EHR** and the chest X-ray images as **CXR**. **(EHR+CXR)_PARTIAL** contains paired and partially paired samples (i.e. samples where chest X-ray is missing). **(EHR + CXR)_PAIRED** contains data samples where both modalities are present. For example, the **(EHR + CXR)_PARTIAL** training set for patient phenotyping contains 8056 samples associated with chest X-rays amongst 42628 samples.

We extract from MIMIC-CXR chest X-ray images and split them based on a random patient split. We then transfer images from the training set to either the validation or test set, in case are were associated with patients in the validation or test splits of the clinical time-series data. This procedure resulted with 325188 images in the training set, 15282 images in the validation set, and 36625 images in the test set. We define $\mathbf{y}_{cxr}$ as a vector of 14 binary radiology labels extracted from radiology reports through CheXpert (Irvin et al., 2019). We denote this uni-modal dataset as **CXR_UNI** and it is fixed across all tasks. We introduce an additional notation for **CXR_PAIRED**, which includes only chest X-ray images within **(EHR+CXR)_PAIRED**, and **EHR_PARTIAL**, which includes only clinical time-series data within **(EHR + CXR)_PARTIAL**.

4.2. Training Strategy with the `MedFuse` Module

The training strategy consists of two steps: pre-training of the modality-specific encoders followed by jointly fine-tuning the encoders and fusion module. During the pre-training stage, we train the image encoder using the full uni-modal training dataset **CXR_UNI** with
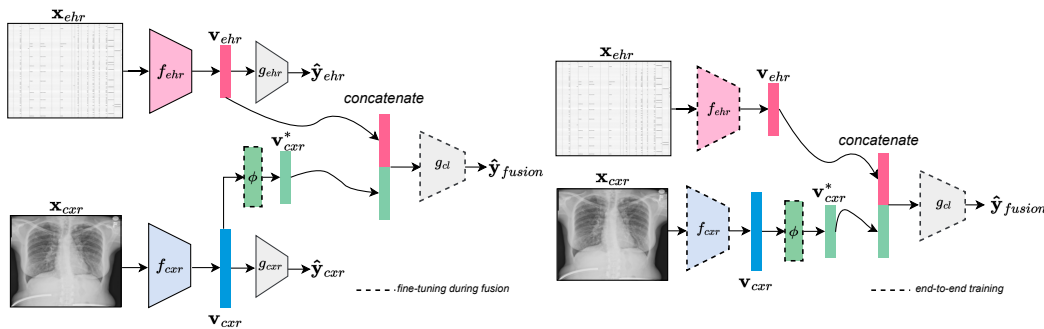
Figure 3: **Architecture of early and joint fusion baselines.** In early fusion (left), the encoders are first pre-trained. Then, we freeze them and fine-tune the projection layer and fusion classification module. In joint fusion (right), the encoders and classification module are randomly initialized and trained end-to-end.

the 14 radiology labels. We also pre-train the clinical time-series data encoder for each task independently using the training sets $\mathbf{EHR_{PARTIAL}}$, since each task is associated with its own set of inputs and labels. After pre-training the modality-specific encoders, we discard the uni-modal classifiers and fine-tune the encoders, projection layer, and `MedFuse` using $(\mathbf{EHR + CXR})_{\mathbf{PARTIAL}}$. We compare this training strategy to fine-tuning the fusion module with randomly initialized feature extractors.

4.3. Baseline Models

We compare the performance of our proposed multi-modal approach to several existing baselines:

- **Early fusion:** The vanilla early fusion approach commonly used in recent work (Huang et al., 2020a) (Figure 3 (left)) assumes the presence of paired data modalities during training and inference. We train two versions. In the first version, we pre-train modality-specific networks independently: f_{cxr} and g_{cxr} with the $\mathbf{CXR_{PAIRED}}$ training set, and f_{ehr} and g_{ehr} with the $\mathbf{EHR_{PAIRED}}$ training set. We then freeze the encoders f_{cxr} and f_{ehr}, concatenate their latent feature representations, and fine-tune a projection layer and a fully connected classification network, denoted as g_{cl} using the $(\mathbf{EHR + CXR})_{\mathbf{PAIRED}}$ training set. In the second version, we use the $(\mathbf{EHR + CXR})_{\mathbf{PARTIAL}}$ training set for fine-tuning the projection layer and g_{cl}. Inspired by Kyono et al. (2021), we learn a vector to substitute for missing chest X-ray images.

- **Joint fusion:** In this setting, we train a network end-to-end including the modality-specific encoders (f_{cxr} and f_{ehr}) and a classification network applied to the concatenated latent representations of the two encoders (Figure 3 (right)). We train two versions. In the first version, we train a randomly initialized network end-to-end using $(\mathbf{EHR + CXR})_{\mathbf{PAIRED}}$. In the second version, we train a randomly initialized network end-to-end using $(\mathbf{EHR + CXR})_{\mathbf{PARTIAL}}$ with a learnable vector to substitute for any missing chest X-ray images.

- **Multi-modal Transfer Module (MMTM):** Originally proposed by Vaezi Joze et al. (2020), this approach also assumes paired input data. We apply an MMTM

Table 1: **Performance results in the uni-modal vs multi-modal setting.** Here, we compare the stacked LSTM network for the clinical time-series data only, with our network using `MedFuse`. In the first four rows, we summarize the AUROC and AUPRC results in the paired setting with uni-modal ($\mathbf{EHR_{PAIRED}}$) and multi-modal data (($\mathbf{EHR + CXR})_{\mathbf{PAIRED}}$). In the last two rows, we show the results on the partially paired test set, with the uni-modal subset ($\mathbf{EHR_{PARTIAL}}$) and multi-modal data (($\mathbf{EHR + CXR})_{\mathbf{PAIRED}}$). All results shown below are for `MedFuse` (OPTIMAL). Best results are shown in bold.

	Modalities		Phenotyping		In-hospital mortality	
Model	Training set	Test set	AUROC	AUPRC	AUROC	AUPRC
LSTM	$\mathbf{EHR_{PAIRED}}$	$\mathbf{EHR_{PAIRED}}$	0.716	0.407	0.818	0.460
			(0.688, 0.743)	(0.367, 0.453)	(0.787, 0.845)	(0.395, 0.535)
LSTM	$\mathbf{EHR_{PARTIAL}}$	$\mathbf{EHR_{PAIRED}}$	0.746	0.453	0.825	0.500
			(0.720, 0.772)	(0.409, 0.502)	(0.793, 0.852)	(0.428, 0.576)
MedFuse	$(\mathbf{EHR + CXR})_{\mathbf{PARTIAL}}$	$\mathbf{EHR_{PAIRED}}$	0.740	0.441	0.833	0.514
			(0.713, 0.767)	(0.398, 0.489)	(0.802, 0.861)	(0.443, 0.584)
MedFuse	$(\mathbf{EHR + CXR})_{\mathbf{PARTIAL}}$	$(\mathbf{EHR + CXR})_{\mathbf{PAIRED}}$	**0.770**	**0.481**	**0.865**	**0.594**
			(0.745, 0.795)	(0.436, 0.531)	(0.837, 0.889)	(0.526, 0.655)
LSTM	$\mathbf{EHR_{PARTIAL}}$	$\mathbf{EHR_{PARTIAL}}$	0.765	0.425	0.861	0.522
			(0.754, 0.777)	(0.404, 0.447)	(0.846, 0.876)	(0.482, 0.564)
MedFuse	$(\mathbf{EHR + CXR})_{\mathbf{PARTIAL}}$	$(\mathbf{EHR + CXR})_{\mathbf{PARTIAL}}$	**0.768**	**0.429**	**0.874**	**0.567**
			(0.756, 0.779)	(0.408, 0.452)	(0.860, 0.888)	(0.529, 0.607)

module after the first LSTM layer in the clinical time-series modality, and either the third or the fourth ResNet layer. We train a randomly initialized network with the MMTM module end-to-end using the $(\mathbf{EHR+CXR})_{\mathbf{PAIRED}}$ training set, and closely follow the training strategy described in the original paper.[2]

- **Dynamic Affine Feature Map Transform (DAFT):** Also requiring paired input data, we use the general purpose DAFT module (Pölsterl et al., 2021) to rescale and shift the feature representations after the first LSTM layer using the chest X-ray representation computed either through the third or fourth layer of ResNet. Similarly, we use $(\mathbf{EHR+CXR})_{\mathbf{PAIRED}}$, and follow the training approach in the original work's respository.[3]

We also compare it with a uni-modal two-layer LSTM network trained with clinical time-series data only, and the method proposed by Hayat et al. (2021a) (Unified) trained with $(\mathbf{EHR + CXR})_{\mathbf{PARTIAL}}$.

4.4. Model Training and Selection

We perform hyperparameter tuning over 10 runs for our proposed network with `MedFuse` and each of the baseline models and their different versions. In each run, we randomly sample a learning rate between 10^{-5} and 10^{-3}, and then choose the model and learning rate that achieve the best AUROC on the respective validation set. For the baselines with architectural choices (i.e. MMTM and DAFT), we choose the architecture that achieves the best performance on the validation set, and report its results on the test set. We use the Adam optimizer (Kingma and Ba, 2014) across all experiments with a batch size of

---

2. https://github.com/haamoon/mmtm
3. https://github.com/ai-med/DAFT/

Table 2: **Performance results on the (EHR + CXR)$_{\textbf{PAIRED}}$ test set.** We show the AUROC and AUPRC results for our proposed approach with `MedFuse` and the baseline models. We include results for early and joint fusion when trained with either (**EHR** + **CXR**)$_{\textbf{PAIRED}}$ or (**EHR** + **CXR**)$_{\textbf{PARTIAL}}$, where the latter uses a learnable vector in the case of a missing chest X-ray image. We also show results of our proposed approach when we fine-tune the fusion module with (**EHR** + **CXR**)$_{\textbf{PARTIAL}}$ and randomly initialized encoders (RI) or pre-trained encoders (PT), and the best version of the latter when using the optimal number of uni-modal samples during fine-tuning (OPTIMAL). Best results are shown in bold.

Task	Phenotyping		In-hospital mortality	
Method	AUROC	AUPRC	AUROC	AUPRC
Early (**EHR** + **CXR**)$_{\textbf{PAIRED}}$	0.753 (0.726, 0.779)	0.453 (0.411, 0.502)	0.827 (0.801, 0.854)	0.485 (0.417, 0.555)
Early (**EHR** + **CXR**)$_{\textbf{PARTIAL}}$	0.739 (0.712, 0.766)	0.435 (0.393, 0.483)	0.818 (0.788, 0.845)	0.467 (0.402, 0.539)
Joint (**EHR** + **CXR**)$_{\textbf{PAIRED}}$	0.747 (0.720, 0.773)	0.446 (0.404, 0.493)	0.825 (0.798, 0.853)	0.506 (0.436, 0.574)
Joint (**EHR** + **CXR**)$_{\textbf{PARTIAL}}$	0.754 (0.727, 0.780)	0.458 (0.415, 0.506)	0.819 (0.785, 0.850)	0.479 (0.413, 0.552)
MMTM (Vaezi Joze et al., 2020)	0.734 (0.707, 0.761)	0.428 (0.387, 0.476)	0.819 (0.788, 0.846)	0.474 (0.402, 0.544)
DAFT (Pölsterl et al., 2021)	0.737 (0.710, 0.764)	0.434 (0.393, 0.482)	0.828 (0.799, 0.854)	0.492 (0.427, 0.572)
Unified (Hayat et al., 2021a)	0.765 (0.742, 0.794)	0.461 (0.417, 0.511)	0.835 (0.808, 0.861)	0.495 (0.424, 0.567)
`MedFuse` (RI)	0.748 (0.721, 0.774)	0.452 (0.408, 0.501)	0.817 (0.785, 0.846)	0.471 (0.404, 0.545)
`MedFuse` (PT)	0.756 (0.729, 0.782)	0.466 (0.420, 0.515)	0.841 (0.813, 0.868)	0.544 (0.477, 0.609)
`MedFuse` (OPTIMAL)	**0.770** (0.745, 0.795)	**0.481** (0.436, 0.531)	**0.865** (0.837, 0.889)	**0.594** (0.526, 0.655)

16. We set the maximum number of epochs to 50 and use early stopping if the validation AUROC does not improve for 15 epochs. We also apply image augmentations as described in Appendix A.1.

With the best learning rate chosen via hyperparameter tuning, we vary the percentage of samples with **EHR** only data in the (**EHR** + **CXR**)$_{\textbf{PARTIAL}}$ training set, fine-tune `MedFuse` accordingly and evaluate it on the validation set. We select the best model based on the best AUROC performance on the (**EHR** + **CXR**)$_{\textbf{PARTIAL}}$ validation set, and report its results on the test set. We denote this chosen model as `MedFuse` (OPTIMAL).

5. Results

In this section, we describe the results for a number of experiments to provide insights on our proposed approach. The learning rates that achieved the best results are summarized in Appendix A.2 for all models. The results on the validation set in the experiments where we vary the percentage of uni-modal samples during training are shown in Appendix A.3. The optimal percentages are 10% for in-hospital mortality prediction, and 20% for phenotype classification.

5.1. Performance Results in the Uni-modal & Multi-modal Settings

In Table 1, we compare our proposed approach to the uni-modal stacked LSTM. As expected, we first observe that the performance of the uni-modal LSTM improves on the **EHR**$_{\textbf{PAIRED}}$ test set, in terms of AUROC and AUPRC for both tasks, when using the larger **EHR**$_{\textbf{PARTIAL}}$ training set. Our proposed approach using `MedFuse` achieves the best performance on the paired test set when the chest X-ray images are used during training and inference as an auxiliary modality (0.770 AUROC and 0.481 AUPRC for phenotype

Table 3: **Performance results on the (EHR + CXR)$_{\mathbf{PARTIAL}}$ test set.** We compare our proposed approach with `MedFuse` with early and joint fusion when trained with (**EHR + CXR**)$_{\mathbf{PARTIAL}}$, including samples with missing chest X-ray images (substituted with a learnable vector). All methods were trained with the full (**EHR + CXR**)$_{\mathbf{PARTIAL}}$ training set, except for `MedFuse` (OPTIMAL) which uses the optimal number of uni-modal samples during fine-tuning. Best results are shown in bold.

Task	Phenotyping		In-hospital mortality	
Method	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
Early	0.748 (0.735, 0.760)	0.394 (0.374, 0.416)	0.860 (0.850, 0.877)	**0.515** (0.477, 0.556)
Joint	0.754 (0.742, 0.766)	0.410 (0.389, 0.433)	0.841 (0.823, 0.857)	0.482 (0.442, 0.525)
MedFuse	**0.758** (0.745, 0.770)	**0.418** (0.396, 0.441)	**0.861** (0.845, 0.874)	0.501 (0.462, 0.543)
MedFuse (OPTIMAL)	**0.768** (0.756, 0.779)	**0.429** (0.408, 0.452)	**0.874** (0.860, 0.888)	**0.567** (0.529, 0.607)

classification, and 0.865 AUROC and 0.594 AUPRC for in-hospital mortality). We note similar, but less significant trends, in the larger partially paired test set, which may be due to the fact that only 18.8% and 26.2% of samples are paired in the phenotyping and in-hospital mortality test sets, respectively.

5.2. Performance Results in the Paired Setting

Since the baseline models were originally designed for paired input, we evaluate all models on the (**EHR + CXR**)$_{\mathbf{PAIRED}}$ test set as shown in Table 2. First, we observe that early fusion and joint fusion perform comparably across both tasks when trained with (**EHR + CXR**)$_{\mathbf{PAIRED}}$, with early fusion achieving a slightly better performance in terms of AUROC. We also note that training early fusion using (**EHR + CXR**)$_{\mathbf{PARTIAL}}$ leads to a drop in AUROC and AUPRC across both tasks, while joint fusion only improves for phenotype classification. Second, we observe that the Unified approach by Hayat et al. (2021a) achieves the best performance amongst all baseline approaches, with 0.765 AU-ROC and 0.461 AUPRC for phenotype classification, and 0.835 AUROC and 0.495 AUPRC for in-hospital mortality prediction. Third, we observe that our proposed approach with `MedFuse` (OPTIMAL) achieves the best performance across both tasks, with 0.770 AUROC and 0.481 AUPRC for phenotype classification, and 0.865 AUROC and 0.594 AUPRC for in-hospital mortality prediction. We also performed an ablation study where we randomly dropped the chest X-ray modality in the paired test set. The results are shown in Appendix A.4. We also compared the use of substituting the missing modality with zeros or with a learnable vector for early and joint fusion and the results are shown in Appendix A.5. The two techniques perform comparably.

5.3. Performance Results in the Partially Paired Setting

In Table 3, we evaluate our proposed approach with `MedFuse` as well as early and joint fusion on the partially paired test set. Compared to early fusion, our proposed approach trained with the full (**EHR + CXR**)$_{\mathbf{PARTIAL}}$ training set achieves a better performance for phenotype classification (0.758 compared to 0.748 AUROC and 0.418 compared to 0.394 AUPRC). It performs comparably with early fusion in the in-hospital mortality prediction
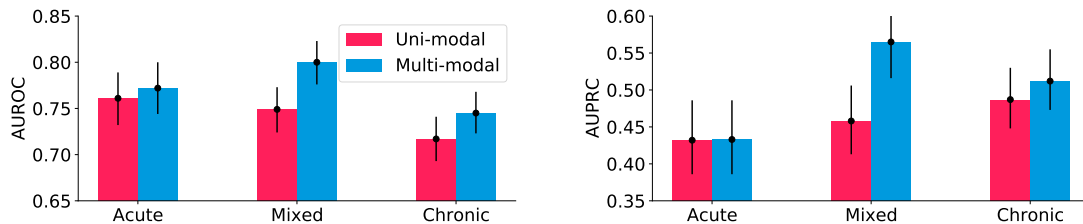
Figure 4: **Performance results across different subsets of labels for the phenotype classification task in the (EHR + CXR)_{PAIRED} test set.** The multi-modal approach with `MedFuse` achieves the highest AUROC and AUPRC gains for the mixed conditions followed by chronic conditions, compared to the results achieved by the uni-modal stacked LSTM with **EHR_{PAIRED}**).

task, although early fusion achieves a better AUPRC. Our approach outperforms joint fusion in the in-hospital mortality setting (0.861 compared to 0.841 AUROC and 0.501 compared to 0.482 AUPRC), and performs comparably for phenotype classification. Overall, `MedFuse` (OPTIMAL), fine-tuned with paired samples and only 10% of uni-modal samples for in-hospital mortality prediction and 20% of uni-modal samples for phenotype classification, achieves the best performance (0.768 AUROC and 0.429 AUPRC for phenotype classification and 0.874 AUROC and 0.567 AUPRC for inhospital mortality prediction). We also performed an ablation study where we varied the percentage of uni-modal samples in the partially paired setting. The results are shown in Appendix A.6.

We also compared the performance of `MedFuse` to an ensemble consisting of (i) `MedFuse` for paired samples, and (ii) a uni-modal LSTM for samples with missing chest X-rays. While the results are comparable, as shown in Appendix A.7, the results imply that an ensemble of strong models may be better suited for some tasks, such as phenotyping. This however requires the training of two models.

5.4. Phenotype-wise Analysis

In Figure 4, we show the AUROC (left) and AUPRC (right) results across different categories of phenotype labels: acute, mixed, and chronic conditions. The label types and their prevalence are listed in Table 4. We note that our approach mostly improves the performance in terms of AUROC and AUROC for mixed and chronic conditions, which are generally hard to predict through uni-modal clinical time-series data (Harutyunyan et al., 2019). In particular, across mixed conditions, the AUROC increases from 0.749 to 0.800, and the AUPRC increases from 0.458 to 0.565. For chronic conditions, the AUROC increases from 0.717 to 0.745 and the AUPRC increases from 0.487 to 0.512. We observe relatively smaller improvements for acute conditions, where the AUROC increases from 0.761 to 0.772 and the AUPRC increases from 0.432 to 0.433. In Table 4, we report the performance across all 25 labels for the paired test set using uni-modal and multi-modal data. We observe an improvement across a number of thorax-related phenotypes, such as pneumonia and pleurisy, which are usually clinically assessed using chest imaging (Long et al., 2017). This further highlights the importance of using the chest X-ray images as auxiliary information along with the clinical time-series data.

Table 4: **Performance results across the different phenotype labels on (EHR + CXR)$_{\text{PAIRED}}$ test set, compared to the uni-modal stacked LSTM with EHR$_{\text{PAIRED}}$.** We report the performance results for the individual phenotypes using AUROC and AUPRC, and show the prevalence of labels in the **(EHR + CXR)$_{\text{PARTIAL}}$** training set, and the **(EHR + CXR)$_{\text{PAIRED}}$** test set. The labels and results in bold indicate that `MedFuse` achieved a performance improvement.

Phenotype	Type	Prevalence Train	Prevalence Test	EHR$_{\text{PAIRED}}$ AUROC	EHR$_{\text{PAIRED}}$ AUPRC	(EHR + CXR)$_{\text{PAIRED}}$ AUROC	(EHR + CXR)$_{\text{PAIRED}}$ AUPRC
Acute and unspecified renal failure	acute	0.269	0.321	0.780 (0.759, 0.800)	0.614 (0.573, 0.655)	0.782 (0.760, 0.802)	**0.618** (0.579, 0.661)
Acute cerebrovascular disease	acute	0.056	0.078	0.903 (0.878, 0.929)	0.501 (0.426, 0.591)	0.888 (0.859, 0.915)	0.496 (0.420, 0.578)
Acute myocardial infarction	acute	0.074	0.093	0.729 (0.694, 0.765)	0.267 (0.218, 0.329)	0.766 (0.732, 0.798)	**0.297** (0.237, 0.361)
Cardiac dysrhythmias	mixed	0.326	0.379	0.664 (0.640, 0.687)	0.552 (0.517, 0.590)	0.708 (0.686, 0.730)	**0.581** (0.543, 0.618)
Chronic kidney disease	chronic	0.206	0.240	0.748 (0.727, 0.771)	0.457 (0.419, 0.505)	0.768 (0.747, 0.789)	**0.485** (0.445, 0.533)
Chronic obstructive pulmonary disease	chronic	0.143	0.148	0.673 (0.640, 0.703)	0.272 (0.231, 0.319)	0.747 (0.721, 0.776)	**0.344** (0.302, 0.398)
Complications of surgical/medical care	acute	0.189	0.226	0.728 (0.703, 0.752)	0.464 (0.420, 0.513)	0.722 (0.698, 0.747)	0.439 (0.395, 0.487)
Conduction disorders	mixed	0.100	0.115	0.719 (0.688, 0.750)	0.252 (0.210, 0.304)	**0.854** (0.822, 0.882)	**0.632** (0.570, 0.692)
Congestive heart failure; nonhypertensive	mixed	0.255	0.295	0.760 (0.738, 0.781)	0.592 (0.553, 0.632)	**0.823** (0.805, 0.843)	**0.679** (0.643, 0.715)
Coronary atherosclerosis and related	chronic	0.311	0.337	0.740 (0.719, 0.763)	0.603 (0.563, 0.643)	0.779 (0.760, 0.799)	**0.631** (0.593, 0.668)
Diabetes mellitus with complications	mixed	0.114	0.120	0.885 (0.866, 0.902)	0.534 (0.473, 0.596)	0.883 (0.862, 0.902)	0.534 (0.473, 0.599)
Diabetes mellitus without complication	chronic	0.172	0.211	0.758 (0.731, 0.781)	0.430 (0.386, 0.481)	0.748 (0.724, 0.772)	0.414 (0.370, 0.463)
Disorders of lipid metabolism	chronic	0.404	0.406	0.689 (0.666, 0.713)	0.598 (0.562, 0.635)	0.707 (0.685, 0.729)	**0.613** (0.577, 0.649)
Essential hypertension	chronic	0.418	0.433	0.678 (0.655, 0.699)	0.617 (0.583, 0.650)	**0.703** (0.682, 0.725)	**0.634** (0.600, 0.667)
Fluid and electrolyte disorders	acute	0.371	0.454	0.737 (0.716, 0.757)	0.696 (0.666, 0.727)	0.733 (0.713, 0.754)	0.687 (0.657, 0.720)
Gastrointestinal hemorrhage	acute	0.070	0.071	0.751 (0.712, 0.785)	0.194 (0.145, 0.254)	0.747 (0.708, 0.783)	0.221 (0.165, 0.287)
Hypertension with complications	chronic	0.215	0.222	0.736 (0.714, 0.758)	0.430 (0.391, 0.475)	**0.764** (0.742, 0.786)	**0.465** (0.421, 0.511)
Other liver diseases	mixed	0.125	0.169	0.716 (0.687, 0.743)	0.359 (0.313, 0.409)	**0.730** (0.704, 0.759)	**0.398** (0.353, 0.450)
Other lower respiratory disease	acute	0.095	0.126	0.610 (0.574, 0.645)	0.194 (0.164, 0.238)	0.599 (0.564, 0.637)	0.176 (0.150, 0.210)
Other upper respiratory disease	acute	0.048	0.054	0.746 (0.692, 0.796)	0.254 (0.185, 0.340)	0.753 (0.705, 0.798)	0.204 (0.148, 0.286)
Pleurisy; pneumothorax; pulmonary collapse	acute	0.067	0.095	0.627 (0.590, 0.661)	0.152 (0.210, 0.121)	0.752 (0.720, 0.783)	**0.212** (0.174, 0.262)
Pneumonia	acute	0.127	0.185	0.765 (0.738, 0.789)	0.416 (0.374, 0.470)	0.790 (0.766, 0.813)	**0.453** (0.407, 0.501)
Respiratory failure; insufficiency; arrest (adult)	acute	0.160	0.282	0.845 (0.827, 0.863)	0.678 (0.637, 0.721)	0.836 (0.817, 0.854)	0.653 (0.612, 0.693)
Septicemia (except in labor)	acute	0.158	0.227	0.813 (0.794, 0.834)	0.572 (0.528, 0.621)	0.809 (0.790, 0.830)	0.564 (0.522, 0.613)
Shock	acute	0.123	0.174	0.865 (0.844, 0.884)	0.617 (0.565, 0.666)	0.864 (0.843, 0.883)	0.604 (0.552, 0.654)
Average	all	-	-	0.746 (0.720, 0.772)	0.453 (0.409, 0.502)	**0.770** (0.745, 0.795)	**0.481** (0.436, 0.531)

Table 5: **Performance of `MedFuse` across different age groups for in-hospital-mortality on the (EHR + CXR)$_{\text{PAIRED}}$) test set, compared to the uni-modal stacked LSTM with EHR$_{\text{PAIRED}}$.** We compare the AUROC and AUPRC for the different age groups. The results in bold indicate improved performance with multi-modal data.

Age group	Positive fraction	EHR$_{\text{PAIRED}}$ AUROC	EHR$_{\text{PAIRED}}$ AUPRC	(EHR+CXR)$_{\text{PAIRED}}$ AUROC	(EHR+CXR)$_{\text{PAIRED}}$ AUPRC
18-40	0.078 (11/141)	0.941 (0.859, 0.988)	0.613 (0.332, 0.883)	0.917 (0.820, 0.980)	**0.521** (0.272, 0.841)
40-60	0.119 (44/369)	0.796 (0.719, 0.865)	0.403 (0.277, 0.541)	**0.822** (0.751, 0.885)	**0.499** (0.360, 0.629)
60-80	0.159 (98/616)	0.846 (0.805, 0.885)	0.576 (0.478, 0.670)	**0.868** (0.830, 0.900)	**0.583** (0.484, 0.678)
> 80	0.227 (56/247)	0.789 (0.722, 0.850)	0.549 (0.428, 0.677)	**0.841** (0.784, 0.891)	**0.616** (0.491, 0.731)

5.5. In-hospital Mortality Age-wise Analysis

We evaluate the performance of our approach across different age groups, as shown in Table 5, and compare it to the uni-modal stacked LSTM. We observe that the AUROC and AUPRC improve across age groups 40-60, 60-80, and >80 years, while the AUROC decreases for the 18-40 years. The latter result needs further investigation with a larger dataset, since the test sets only contain 11 positive samples for the youngest age group. Additionally, there are variations in the relative improvements. For example, the AUPRC increases by 24% for the 40-60 years group, compared to 1.3% in the 60-80 years group.

6. Discussion

In this paper, we present a multi-modal fusion approach, named `MedFuse`, and new benchmark results for integrating partially paired clinical time-series data and chest X-ray images. We evaluate it for two popular benchmark tasks, namely in-hospital mortality prediction and phenotype classification, using publicly available datasets MIMIC-IV and MIMIC-CXR.

Our study has several strengths. First, our approach is simple and easy to implement. The results show that the proposed approach performs better than the uni-modal LSTM baseline, as it considers chest X-ray images, when available, as an additional source of information. In addition, the approach outperforms several baselines, and the phenotype-wise and age-wise analysis provide some insight as to where it improves performance. We conclude that the proposed method is overall a better choice than the baseline methods because (i) the LSTM-based fusion module can inherently deal with missingness (i.e., partially paired data), and (ii) the combination of the architecture and the training procedure provides performance gains. Otherwise, the size of the partially paired training set does not seem to be correlated with the performance improvements, as illustrated with the validation set results in Appendix A.3. The results overall highlight the promise of multi-modal fusion in improving the performance of clinical prediction models. Multi-modal learning is also generally more closely aligned with the decision-making process of clinicians, who consider multiple sources of information when assessing a patient.

Moreover, in contrast with conventional multi-modal approaches that assume paired input, our proposed method is more flexible since it can process samples with missing chest X-ray images. There is a rising interest in learning cross-modal interactions between modalities during training time and in reconstructing missing modalities (Ngiam et al., 2011; Xin et al., 2021; Ma et al., 2021; Sylvain et al., 2021; Ma et al., 2021). In contrast with natural multi-modal datasets, assuming a high degree of correlation in such settings is not a trivial task in healthcare especially when the modalities do not necessarily share the same labels, and this is an area of future work. The difficulty stems from the sparse and asynchronous nature of medical data, i.e. it would be difficult to use a biopsy report for skin tissue to reconstruct common thorax diseases features (Hayat et al., 2021a). Additionally, some of the existing work for learning cross-modal interactions assumes the presence of all modalities during training (Sylvain et al., 2021).

Another strength is that the approach can be easily scaled to more than two modalities with no amendments to the fusion loss function, compared to existing work where the complexity of the computation increases with the number of modalities (Hayat et al., 2021a). However, this requires evaluation and is an area of future work. We also do not assume any correlation among the input modalities, in terms of information content or assigned labels.

Furthermore, we formalize and introduce new benchmark results for two popular tasks that are typically evaluated in the context of clinical time-series data only (Harutyunyan et al., 2019). By gaining access to the MIMIC-IV and MIMIC-CXR datasets (Johnson et al., 2021, 2019), researchers can utilize our open-access data pre-processing pipeline and introduce new results for direct comparison.

Limitations. The study also has its own limitations. To begin with, we focus on tasks pertaining to the integration of clinical time-series data and chest X-ray images from a single data source, and we evaluate our work on two benchmark tasks due to limited resources.

The original work by Harutyunyan et al. (2019) includes two other tasks, decompensation prediction and length of stay prediction, which we would like to evaluate our method on in the future. We also do not run any experiments on settings where the clinical time-series data may be missing, but the chest X-ray image is available. In future work, this requires the definition of additional benchmark tasks where the chest X-ray image is the primary modality. Since we currently evaluate our method with two input modalities only, another interesting next step would be to use more than two to further evaluate the robustness of the model, considering its scalability. In its current formulation, the model also lacks interpretability, since we mainly focus on fusion within the scope of this paper. We later plan to explore incorporating attention layers (Vaswani et al., 2017) at the input level of the feature encoders to evaluate the importance of features within each modality, and within the fusion module to evaluate the overall informativeness of each modality. On a related note, our work can benefit from performing instance-level analysis. However, this requires clinical expertise that bridges between chest X-ray image and clinical time-series analysis, which we are currently missing. To realize the full potential of multi-modal learning, there is more work to be done to understand the clinical underpinnings of multi-modal fusion. Overall, the study highlights an extremely worthwhile direction to further leverage the value of multi-modal learning in healthcare, especially as the diversity and quantity of medical data continues to increase.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

Hiba Asri, Hajar Mousannif, Hassan Al Moatassime, and Thomas Noel. Big data in healthcare: Challenges and opportunities. In *2015 International Conference on Cloud Technologies and Applications (CloudTech)*, pages 1–7. IEEE, 2015.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In

2015 IEEE International conference on image processing (ICIP), pages 168–172. IEEE, 2015.

Yi-Wen Chen, Yi-Hsuan Tsai, and Ming-Hsuan Yang. End-to-end multi-modal video temporal grounding. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

Declan Grant, Bartłomiej W. Papież, Guy Parsons, Lionel Tarassenko, and Adam Mahdi. Deep learning classification of cardiomegaly using combined imaging and non-imaging icu data. In *Medical Image Understanding and Analysis*, pages 547–558, Cham, 2021. Springer International Publishing.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96, 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0103-9.

Nasir Hayat, Munawar Hayat, Shafin Rahman, Salman Khan, Syed Waqas Zamir, and Fahad Shahbaz Khan. Synthesizing the unseen for zero-shot object detection. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020.

Nasir Hayat, Krzysztof J. Geras, and Farah E. Shamout. Towards dynamic multi-modal phenotyping using chest radiographs and physiological data, 2021a.

Nasir Hayat, Hazem Lashen, and Farah E. Shamout. Multi-label generalized zero shot learning for the classification of disease in chest radiographs. *Proceedings of the 6th Machine Learning for Healthcare Conference, PMLR*, 2021b.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Haithem Hermessi, Olfa Mourali, and Ezzeddine Zagrouba. Multimodal medical image fusion review: Theoretical background and recent advances. *Signal Processing*, 183:108036, 2021. ISSN 0165-1684. doi: https://doi.org/10.1016/j.sigpro.2021.108036.

Danliang Ho, Iain Tan, and Mehul Motani. Predictive models for colorectal cancer recurrence using multi-modal healthcare data. *Proceedings of the Conference on Health, Inference, and Learning. 2021*, pages 204–213, 2021.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P. Lungren. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digital Medicine*, 3(1):136, Oct 2020a. ISSN 2398-6352. doi: 10.1038/s41746-020-00341-z.

Shih-Cheng Huang, Anuj Pareek, Roham Zamanian, Imon Banerjee, and Matthew P. Lungren. 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, Dec 2020b. ISSN 2045-2322. doi: 10.1038/s41598-020-78888-w.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

Alex Pappachen James and Belur V. Dasarathy. Medical image fusion: A survey of the state of the art. *Information Fusion*, 19:4–19, 2014. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2013.12.002. Special Issue on Information Fusion in Medical Image Computing and Systems.

Zhicheng Jiao, Ji Whae Choi, Kasey Halsey, Thi My Linh Tran, Ben Hsieh, Dongcui Wang, Feyisope Eweje, Robin Wang, Ken Chang, Jing Wu, Scott A. Collins, Thomas Y. Yi, Andrew T. Delworth, Tao Liu, Terrance T. Healey, Shaolei Lu, Jianxin Wang, Xue Feng, Michael K. Atalay, Li Yang, Michael Feldman, Paul J. L. Zhang, Wei-Hua Liao, Yong Fan, and Harrison X. Bai. Prognostication of patients with covid-19 using artificial intelligence based on chest x-rays and clinical data: a retrospective study. *The Lancet Digital Health*, 3(5):e286–e294, May 2021. ISSN 2589-7500. doi: 10.1016/S2589-7500(21)00039-X.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. "mimic-iv" (version 1.0). physionet (2021). https://doi.org/10.13026/s6n6-xd98., 2021.

Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Trent Kyono, Yao Zhang, Alexis Bellot, and Mihaela van der Schaar. Miracle: Causally-aware imputation via learning missing data mechanisms. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Garam Lee, Byungkon Kang, Kwangsik Nho, Kyung-Ah Sohn, and Dokyoon Kim. Mildint: Deep learning-based multimodal longitudinal data integration framework. *Frontiers in Genetics*, 10:617, 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00617.

Y. Li, H. Wang, and Y. Luo. A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1999–2004, Los Alamitos, CA, USA, dec 2020. IEEE Computer Society.

Yi Li, Junli Zhao, Zhihan Lv, and Jinhua Li. Medical image fusion method by deep learning. *International Journal of Cognitive Computing in Engineering*, 2:21–29, 2021. ISSN 2666-3074.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

Rahul Lohan. Imaging of icu patients. *Thoracic Imaging: Basic to Advanced*, pages 173–194, Jan 2019. doi: 10.1007/978-981-13-2544-1_7.

Brit Long, Drew Long, and Alex Koyfman. Emergency medicine evaluation of community-acquired pneumonia: history, examination, imaging and laboratory assessment, and risk scores. *The Journal of emergency medicine*, 53(5):642–652, 2017.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. *arXiv preprint arXiv:2103.05677*, 2021.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, 2011.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. 2019.

Angshuman Paul, Thomas C Shen, Sungwon Lee, Niranjan Balachandar, Yifan Peng, Zhiyong Lu, and Ronald M Summers. Generalized zero-shot chest x-ray diagnosis through trait-guided multi-view semantic embedding with self-training. *IEEE transactions on medical imaging*, PP, February 2021. ISSN 0278-0062. doi: 10.1109/tmi.2021.3054817.

Sebastian Pölsterl, Tom Nuno Wolf, and Christian Wachinger. Combining 3D Image and Tabular Data via the Dynamic Affine Feature Map Transform. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 688–698, 2021. doi: 10.1007/978-3-030-87240-3_66.

Esther Puyol-Antón, Baldeep S Sidhu, Justin Gould, Bradley Porter, Mark K Elliott, Vishal Mehta, Christopher A Rinaldi, and Andrew P King. A multimodal deep learning model for cardiac resynchronisation therapy response prediction. *arXiv preprint arXiv:2107.10662*, 2021.

Farah E. Shamout, Yiqiu Shen, Nan Wu, Aakash Kaku, Jungkyu Park, Taro Makino, Stanisław Jastrzebski, Jan Witowski, Duo Wang, Ben Zhang, Siddhant Dogra, Meng Cao, Narges Razavian, David Kudlowitz, Lea Azour, William Moore, Yvonne W. Lui, Yindalon Aphinyanaphongs, Carlos Fernandez-Granda, and Krzysztof J. Geras. An artificial intelligence system for predicting the deterioration of covid-19 patients in the emergency department. *npj Digital Medicine*, 4(1):80, May 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00453-0.

Wei Shao, Tongxin Wang, Liang Sun, Tianhan Dong, Zhi Han, Zhi Huang, Jie Zhang, Daoqiang Zhang, and Kun Huang. Multi-task multi-modal learning for joint diagnosis and prognosis of human cancers. *Medical Image Analysis*, 65:101795, 2020. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2020.101795.

Dhruv Sharma, Sanjay Purushotham, and Chandan K. Reddy. Medfusenet: an attention-based multimodal deep learning model for visual question answering in the medical domain. *Scientific Reports*, 11(1):19826, Oct 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-98390-1.

Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, 2017.

Tom Sonsbeek and Marcel Worring. *Towards Automated Diagnosis with Attentive Multimodal Learning Using Electronic Health Records and Chest X-Rays*, pages 106–114. 10 2020. ISBN 978-3-030-60945-0. doi: 10.1007/978-3-030-60946-7_11.

Tristan Sylvain, Francis Dutil, Tess Berthier, Lisa Di Jorio, Margaux Luck, Devon Hjelm, and Yoshua Bengio. Cmim: Cross-modal information maximization for medical imaging. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1190–1194. IEEE, 2021.

Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L. Iuzzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Luís A. Vale-Silva and Karl Rohr. Multisurv: Long-term cancer survival prediction using multimodal deep learning. *medRxiv*, 2020. doi: 10.1101/2020.08.06.20169698.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

WHO et al. International classification of diseases ninth revision (icd-9). 63(45):343–344, 1988.

Mingxiang Wu, Xiaoling Zhong, Quanzhou Peng, Mei Xu, Shelei Huang, Jialin Yuan, Jie Ma, and Tao Tan. Prediction of molecular subtypes of breast cancer using bi-rads features based on a "white box" machine learning approach in a multi-modal imaging setting. *European Journal of Radiology*, 114, 03 2019. doi: 10.1016/j.ejrad.2019.03.015.

Bowen Xin, Jing Huang, Yun Zhou, Jie Lu, and Xiuying Wang. Interpretation on deep multimodal fusion for diagnostic classification. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2021. doi: 10.1109/IJCNN52387.2021.9534148.

Tao Xu, Han Zhang, Xiaolei Huang, Shaoting Zhang, and Dimitris N Metaxas. Multimodal deep learning for cervical dysplasia diagnosis. In *International conference on medical image computing and computer-assisted intervention*, pages 115–123. Springer, 2016.

Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. Multimodal transformer with multi-view visual representation for image captioning. *IEEE transactions on circuits and systems for video technology*, 30(12):4467–4480, 2019.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34, 2021.

Tianyu Zhang, Luyi Han, Yuan Gao, Xin Wang, Regina Beets-Tan, and Ritse Mann. Predicting molecular subtypes of breast cancer using multimodal deep learning and incorporation of the attention mechanism. In *Medical Imaging with Deep Learning*, 2021a.

Yizhen Zhang, Minkyu Choi, Kuan Han, and Zhongming Liu. Explainable semantic space by grounding language to vision with cross-modal contrastive learning. In *Advances in Neural Information Processing Systems*, 2021b.

Appendix A.

A.1. Image Augmentations

For the chest X-ray images, we apply a series of transformations during pre-training and fine-tuning across all experiments and tasks. Specifically, we resize each image to 256×256 pixels, randomly apply a horizontal flip, and apply a set of random affine transformations, such as rotation, scaling, shearing, and translation. We then apply a random crop to obtain an image of size 224×224 pixels. During validation and testing, we perform image resizing to 256×256 and apply a center crop to 224×224 pixels.

A.2. Hyperparameter Search Results

The results of hyperparameter tuning are shown in Table A1. We summarize the learning rates that achieved the best performance for each model.

Table A1: **Learning rates that achieved the best results during hyperparameter search.** We conducted 10 runs for each model with randomly sampled learning rates between 10^{-5} and 10^{-3}. For MMTM and DAFT, we additionally selected the version that achieved the best validation set AUROC.

Task	Phenotyping	In-hospital mortality
Method	**Learning rate**	
LSTM trained with $\mathbf{EHR_{PAIRED}}$	8.866×10^{-5}	1.000×10^{-4}
LSTM trained with $\mathbf{EHR_{PARTIAL}}$	5.399×10^{-4}	5.399×10^{-4}
Early trained with $\mathbf{(EHR + CXR)_{PARTIAL}}$	9.084×10^{-5}	3.095×10^{-4}
Early trained with $\mathbf{(EHR + CXR)_{PAIRED}}$	3.833×10^{-5}	9.515×10^{-5}
Joint trained with $\mathbf{(EHR + CXR)_{PARTIAL}}$	3.831×10^{-5}	7.565×10^{-4}
Joint trained with $\mathbf{(EHR + CXR)_{PAIRED}}$	5.652×10^{-5}	4.032×10^{-5}
MMTM*	5.326×10^{-5}	4.355×10^{-5}
DAFT**	6.493×10^{-5}	6.493×10^{-5}
Unified	2.042×10^{-4}	2.606×10^{-4}
`MedFuse` (Randomly initialized encoders)	4.741×10^{-5}	9.382×10^{-5}
`MedFuse` (Pre-trained encoders)	7.347×10^{-5}	1.452×10^{-5}

*We trained two versions of MMTM for each task, where the MMTM module is placed after the third or fourth ResNet layer. Placing it after the fourth layer achieved the best performance for both tasks.

**We trained two versions of DAFT for each task, where we transform the LSTM representation either after the third or fourth ResNet layer. Placing it after the third layer achieved the best performance for phenotype classification, whereas placing it after the fourth layer achieved the best performance for in-hospital mortality.

A.3. Percentage of Uni-modal Samples within the Training Set

We also run experiments where we vary the percentage of uni-modal samples during fine-tuning. The best AUROC results for both tasks on the validation set are shown in Figure A1. For in-hospital mortality (shown in red), we notice that a relatively smaller portion of uni-modal samples (10%) achieves the best performance. For patient phenotyping (shown in blue), we observe a similar trend where the best AUROC is achieved with only 20% of uni-modal samples. We fix the sampling percentage that achieves the best validation AUROC across all experiments, unless noted otherwise. Hence, this highlights that the

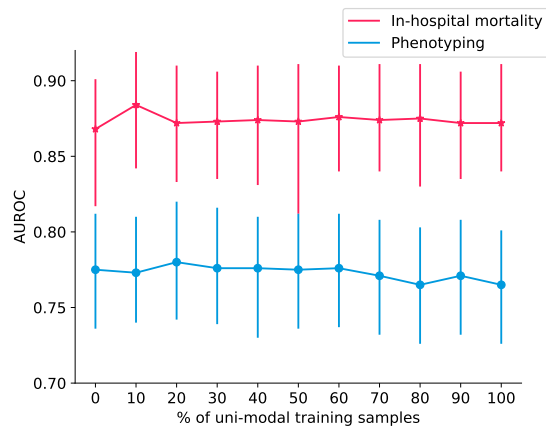best performance gains of `MedFuse` are achieved even with a small percentage of uni-modal samples.

Figure A1: **Performance on the validation set when varying the sampling percentage for uni-modal training samples.** The plot shows the AUROC on the validation set for different percentages of randomly selected uni-modal training samples.

A.4. Percentage of Uni-modal Samples within the Paired Test Set

We also performed an ablation study where we randomly dropped the chest X-ray modality for a percentage of samples in the paired test set. The results are shown in Figure A2. We observe that the as the percentage of dropping increases, the AUROC decreases for both tasks.
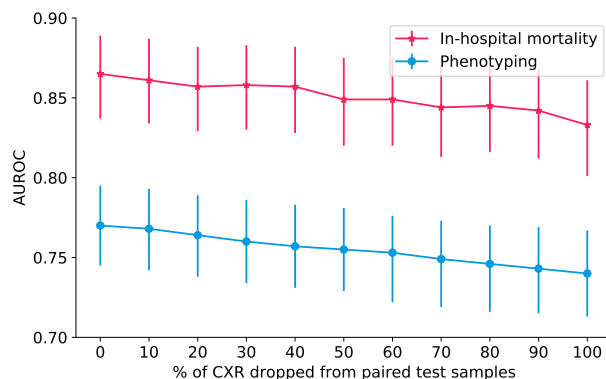

Figure A2: **Performance on the test set with randomly dropped CXR modality in the paired test set.** The plot shows the AUROC on the paired test set for different percentages of randomly dropped CXR modality from paired test samples.

A.5. Missing Modality with Early and Joint Fusion

We also ran initial experiments to compare the learnable vector with imputing zeros for a missing chest X-ray modality. The results are shown in Table A2. We note that the results are comparable with no obvious differences.

Table A2: **Missing modality with early and joint fusion.** We report the AUROC and AUPRC results on the entire test set ($\textbf{EHR}_{\textbf{PARTIAL}}$), including samples with missing chest X-ray images (substituted with a zeros or a learnable vector) All methods below were pre-trained using the ($\textbf{EHR}_{\textbf{PARTIAL}}$) training set and a fixed learning rate of 0.0001.

	Task	Phenotyping		In-hospital mortality	
Method	**Missing Vector**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
Joint	Zeros	0.756	0.406	0.843	0.466
Joint	Learnable	0.752	0.402	0.853	0.486
Early	Zeros	0.743	0.392	0.842	0.481
Early	Learnable	0.742	0.388	0.851	0.489

A.6. Percentage of Uni-modal Samples within the Partially Paired Test Set

We performed another ablation study where we varied the number of uni-modal samples in the partially paired test set. The results are shown in Figure A3. Hence, including 0% of uni-modal test samples is equivalent to the fully paired test set. We observe an increase in the AUROC in the in-hospital mortality task, as the percentage of uni-modal samples increase, but a more a consistent AUROC in the phenotyping task. We do however observe that the widths of the confidence intervals decrease as the percentage of uni-modal samples increases across both tasks.
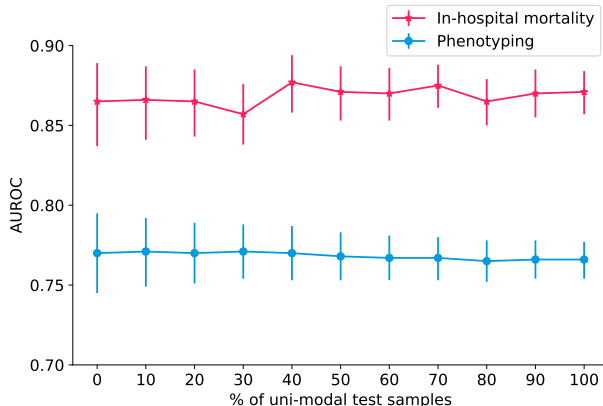

Figure A3: **Performance on the test set when varying the percentage of uni-modal samples.** The plot shows the AUROC on the partial test set for different percentages of randomly selected uni-modal samples.

A.7. Ensemble of uni-modal and multi-modal models

We ran another experiment to compare the performance of `MedFuse` to that of an ensemble of two models: an EHR only model that computes predictions for partial input (i.e., not associated with a chest X-ray) using LSTM, and a paired model that computes predictions for paired input using `MedFuse`. The results are shown in Table A3. We observe that the ensemble slightly outperforms `MedFuse` for phenotyping only. This implies that an ensemble of strong models may be better suited for some tasks, such as phenotyping, which however requires the training of two models.

Table A3: **MedFuse compared to an ensemble evaluation.** We report the AUROC and AUPRC results on the partially paired test set.

Task	Phenotyping		In-hospital mortality	
Method	AUROC	AUPRC	AUROC	AUPRC
Ensemble	0.770 (0.759, 0.782)	0.431 (0.410, 0.454)	0.870 (0.857, 0.884)	0.547 (0.509, 0.589)
MedFuse	0.768 (0.756, 0.779)	0.429 (0.408, 0.452)	0.874 (0.860, 0.888)	0.567 (0.529, 0.607)