MedMod: Multimodal Benchmark for Medical Prediction Tasks with Electronic Health Records and Chest X-Ray Scans

Shaza Elsharief [1] SHAZA.ELSHARIEF@NYU.EDU
Saeed Shurrab [1,2] SAEED.SHURRAB@NYU.EDU
Baraa Al Jorf [1,2] BARAA.AL.JORF@NYU.EDU
L. Julián Lechuga López [1,2] LEOPOLDO.LECHUGA@NYU.EDU
Krzysztof J. Geras [2] K.J.GERAS@NYU.EDU
Farah E. Shamout [1,2] FARAH.SHAMOUT@NYU.EDU
[1] *New York University Abu Dhabi, Abu Dhabi, UAE*
[2] *New York University, New York, USA*

Abstract

Multimodal machine learning provides a myriad of opportunities for developing models that integrate multiple modalities and mimic decision-making in the real-world, such as in medical settings. However, benchmarks involving multimodal medical data are scarce, especially routinely collected modalities such as Electronic Health Records (EHR) and Chest X-ray images (CXR). To contribute towards advancing multimodal learning in tackling real-world prediction tasks, we present MedMod, a multimodal medical benchmark with EHR and CXR using publicly available datasets MIMIC-IV and MIMIC-CXR, respectively. MedMod comprises five clinical prediction tasks: clinical conditions, in-hospital mortality, decompensation, length of stay, and radiological findings. We extensively evaluate multimodal supervised learning models and self-supervised learning frameworks, making our code and models open-source.

Data and Code Availability The data used for this study is available from the public MIMIC-IV database (Johnson et al., 2023) and MIMIC-CXR database (Johnson et al., 2019). To facilitate the use of our benchmark and pre-trained models as feature extractors by the research community, we make all of the code and models open source at: https://github.com/nyuad-cai/MedMod. We also create a public leaderboard to support future research at: https://github.com/nyuad-cai/medmodleaderboard.

Institutional Review Board (IRB) This work did not involve human subjects, so IRB approval was not required.

1. Introduction

Multimodal learning involves leveraging multiple sources of information to build models with a better understanding and representation of real-world data. Multimodal fusion models aggregate information from multiple modalities with the aim of improving predictive performance in downstream tasks (Baltrušaitis et al., 2018; Ngiam et al., 2011). Despite their success, many state-of-the-art models are evaluated on artificially standardized multimodal datasets that do not reflect the complexity and variability of real-world data, such as hateful memes (Kiela et al., 2020) or colored MNIST (Arjovsky et al., 2019). Hence, there is a critical need for new multimodal benchmarks to assess the generalizability of multimodal learning in practical real-world use-cases.

While fields such as computer vision and natural language processing have access to numerous large multimodal datasets, healthcare lacks similarly comprehensive benchmarks. Considering the multimodal nature of clinical decision-making (Kline et al., 2022) and the wide availability of multimodal data across healthcare institutions, such as Electronic Health Records (EHR), medical imaging, and clinical notes, multimodal learning holds a lot of promise for improving medical prediction tasks (Kline et al., 2022; Soenksen et al., 2022; Amal et al., 2022). Data extracted from the patient's EHR is intrinsically multimodal, as it includes all relevant patient data, such as medical history, diagnoses, vital signs, lab results, treatments plans, and administered medication (Shickel et al., 2017; Pivovarov et al., 2015), making it an invaluable source of contextual information for understanding patient status. Further-

more, EHR data has the potential to be combined with other modalities to enable multimodal learning with varying sources of data. The rise of multimodal learning involving EHR was especially evident during the COVID-19 pandemic (Shamout et al., 2021; Satterfield et al., 2021; Estiri et al., 2021) due to the increased accessibility of publicly available Chest X-Ray (CXR) images. Chest radiography is considered to be a low-cost and widely used modality globally compared to other modalities such as computed tomography and magnetic resonance imaging. The integration of both EHR and CXR data for clinical prediction tasks is promising not only for its expected clinical impact, but also for enabling the evaluation and development of more sophisticated methodologies. These methodologies aim to capture complex interactions within multimodal data, leading to improved predictive performance (Zheng et al., 2022; Bardak and Tan, 2021; El-Sappagh et al., 2021).

To address the aforementioned challenges and the lack of comprehensive medical benchmarks, in this paper we present `MedMod`, a multimodal benchmark for clinical prediction tasks using EHR and CXR data. An overview of the benchmark is shown in Figure 1. The goal is to establish a suite of benchmark tasks with routine clinical features and medical imaging, with a focus on clinical outcome prediction tasks. The `MedMod` benchmark is designed for both supervised and self-supervised learning applications. For supervised learning, we establish a suite of tasks relevant to acute care for benchmarking of multimodal fusion approaches. For self-supervised learning, we aim to provide a basis for developing methods that are able to learn multimodal representations which are agnostic to the downstream task. The development of pretrained feature extractors without labels is highly relevant, as it means that they can be applied to situations where annotated data is not largely available. Hence, we believe there is merit in including both of these machine learning paradigms in our benchmark. To the best of our knowledge, `MedMod` is the first comprehensive EHR and CXR benchmark developed for five clinical prediction tasks of high relevance in acute care, including state-of-the-art supervised and self-supervised learning models.

Our main contributions are summarized as follows:

- We propose a diverse and comprehensive multimodal clinical benchmark (`MedMod`) using two publicly available datasets MIMIC-IV (Johnson et al., 2023) and MIMIC-CXR (Johnson et al., 2019) and extending the work of Harutyunyan et al. (2019), comprising five clinical prediction tasks: in-hospital mortality, prediction of clinical conditions, decompensation, length of stay, and radiological findings.
- We perform extensive evaluations of six supervised learning models, encompassing vanilla fusion paradigms (early, joint, late) and other sophisticated multimodal frameworks. We also evaluate three state-of-the-art self-supervised learning methods and present a unified evaluation scheme and protocol for each task.
- We publicly release our code, models, and implementation for all of the proposed clinical prediction tasks to enhance the usability of `MedMod` by the research community and support the advancement of multimodal learning. We also introduce a public leaderboard to support future work.

2. Related Work

Recently, there have been numerous research efforts towards developing benchmarks for medical prediction tasks. Such benchmarks are essential for advancing machine learning research in healthcare (Xie et al., 2022; Strodthoff et al., 2020). They enable standardizing evaluation as well as comparing different machine learning methods and facilitating their reproducibility (Dueben et al., 2022; Pereira et al., 2024). Several studies introduced EHR-based benchmarks covering a wide range of tasks, such as prediction of mortality, length of stay, and patient diagnosis (Harutyunyan et al., 2019; McDermott et al., 2021; Wornow et al., 2024; Gao et al., 2024). Similarly, there are many publicly available medical imaging datasets that are suitable for various tasks, such as disease detection and segmentation (Phillips et al., 2020; Holste et al., 2023; Chen et al., 2023; Kermany et al., 2018; Ji et al., 2022).

Although there are many unimodal benchmarks based on EHR (Harutyunyan et al., 2019; McDermott et al., 2021; Wornow et al., 2024; Gao et al., 2024) and CXR (Phillips et al., 2020; Holste et al., 2023; Chen et al., 2023) data independently, multimodal benchmarks that integrate both modalities are still relatively scarce (Heiliger et al., 2023; Poon, 2023). This scarcity is attributed to both the lack of publicly available datasets and the difficulty of collecting multiple and diverse medical data modalities for the same group of patients (Shaik et al., 2023). For example,

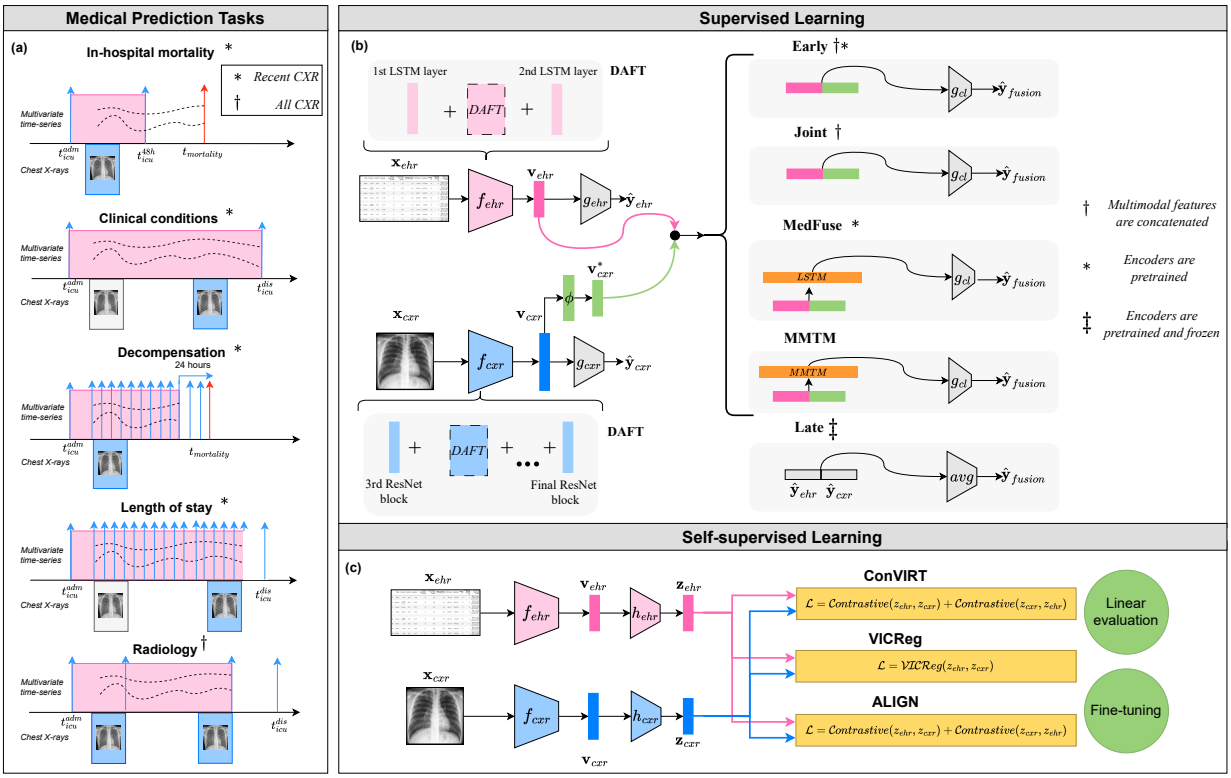

Figure 1: **Overview of the MedMod benchmark.** The main components of the MedMod benchmark include the medical prediction tasks and the training schemes. (a) The MedMod benchmark involves five medical prediction tasks including in-hospital mortality prediction, clinical conditions prediction, decompensation prediction, length of stay prediction, and radiological findings classification. t_{icu}^{adm} represents the time of admission to the ICU, t_{icu}^{dis} represents the time of discharge from the ICU, the red arrow indicates the time of death in the ICU, and t_{icu}^{48h} represents the 48 hour mark at which the prediction is made for the in-hospital mortality task. Note that only the CXRs shaded in blue and EHRs shaded in pink are used as input to the model. For a detailed description of each task set-up, refer to Section 4. (b) Illustration of the supervised learning methods including early, joint, and late fusion as vanilla fusion paradigms, and MedFuse, MMTM, and DAFT as sophisticated fusion models. **Early Fusion** directly concatenates the encoder outputs $\mathbf{v}_{\text{ehr}}$ and $\mathbf{v}_{\text{cxr}}$ (i.e., $[\mathbf{v}_{\text{ehr}}; \mathbf{v}_{\text{cxr}}]$) from randomly initialized encoders, while **Joint Fusion** also concatenates $[\mathbf{v}_{\text{ehr}}; \mathbf{v}_{\text{cxr}}]$ but learns both f_{ehr} and f_{cxr} entirely from scratch in a shared training process. In contrast, **Late Fusion** first obtains unimodal predictions $\hat{y}_{\text{ehr}}$ and $\hat{y}_{\text{cxr}}$ from frozen pretrained encoders then computes the final output by averaging $\hat{y}_{\text{fusion}} = \frac{1}{2}(\hat{y}_{\text{ehr}} + \hat{y}_{\text{cxr}})$. The more advanced fusion models - **MedFuse**, **MMTM**, and **DAFT** - introduce additional layers or modules for richer multimodal interactions. **MedFuse** relies on pretrained encoders f_{ehr} and f_{cxr}, concatenates the resulting features $[\mathbf{v}_{\text{ehr}}; \mathbf{v}_{\text{cxr}}]$, and then passes them through an LSTM to generate the prediction $\hat{y}_{\text{fusion}}$. **MMTM** inserts specialized modules that exchange channel-level features between the latent representations $\mathbf{v}_{\text{ehr}}$ and $\mathbf{v}_{\text{cxr}}$ to better capture cross-modal interactions. Finally, **DAFT**, as illustrated in the dashed boxes within the LSTM and ResNet pipelines, introduces dynamic affine transformations that allow $\mathbf{v}_{\text{ehr}}$ and $\mathbf{v}_{\text{cxr}}$ to modulate each other's feature extraction stages. (c) Illustration of the self-supervised pretraining frameworks, including ConVIRT, VICReg, and ALIGN, and the evaluation protocols used (linear evaluation and fine-tuning). The EHR and CXR input are processed through encoders f_{ehr} and f_{cxr} followed by projection heads h_{ehr} and h_{cxr} to produce the embeddings z_{ehr} and z_{cxr} on which the loss is computed. **ConVIRT** and **ALIGN** employ a bi-directional contrastive loss, while **VICReg** uses the three-term VICReg loss. For a detailed description of each of the models, refer to Section 3.

patients may not always require a CXR scan during their hospital visits. Moreover, for patients who do have both modalities collected within the same hospital encounter, pairing the two modalities requires additional pre-processing steps and domain expertise. This further hinders the development of multimodal clinical benchmarks (Krones et al., 2024). For instance, the Medical Information Mart for Intensive Care (MIMIC)-IV (Johnson et al., 2023) and MIMIC-CXR (Johnson et al., 2019) datasets include EHR and CXR data, respectively, and overlap with respect to the included patients. However, linking both datasets is not straightforward (Wornow et al., 2024). This is attributed to the remarkable difference between both modalities in terms of the source of each modality and the collection scheme, which leads to discrepancies in timing, granularity, and relevance. In addition, the fact that not all patients may have corresponding data for all modalities leads to incomplete datasets that can affect the reliability of the multimodal tasks. Lastly, the privacy regulations in healthcare impose further challenges on gathering diverse multimodal datasets suitable for benchmarking (Rajpurkar et al., 2022). As such, these challenges emphasize the need for comprehensive multimodal benchmarks that reflect the complexity and diversity of real-world clinical data.

Due to the increase in the availability of multimodal medical datasets (Demner-Fushman et al., 2016; Ionescu et al., 2018; Bustos et al., 2020; Littlejohns et al., 2020; Clark et al., 2013), several benchmarks have recently emerged. We provide an overview of benchmarks that were recently introduced and are closely related to the scope of our study in Table 1. While all included benchmarks use EHR as an input modality, only a few consider medical imaging. In the context of acute care, only one study introduced a multimodal fusion benchmark (supervised learning) for the in-hospital mortality and clinical conditions classification tasks using CXR and EHR (Hayat et al., 2022). Two studies included both medical imaging and EHR for pulmonary embolism detection (Zhou et al., 2021; Huang et al., 2023), and another for question answering (Bae et al., 2024). A limited number of benchmarks included self-supervised learning models, specifically for EHR time-series classification (McDermott et al., 2021), electrocardiogram question answering (Oh et al., 2024), and EHR for in-hospital prediction tasks.

3. Methods

3.1. Preliminaries

To present the benchmark and its development process, we first introduce some relevant notation. For a given patient p, let $\mathbf{x}_{ehr} \in \mathbb{R}^{d \times t}$ be a multivariate time-series modality consisting of d features and t time steps representing an Intensive Care Unit (ICU) stay. Also, let $\mathbf{x}^i_{cxr} \in \mathbb{R}^{h \times w \times c}$ be the i-th CXR scan collected from patient p during the same ICU stay, where h, w, and c represent the image height, width, and number of channels, respectively. Our goal is to build a set of medical prediction tasks suitable for multimodal benchmarking in both supervised and self-supervised learning settings. Each supervised task is associated with its specific ground-truth label set y for a given unimodal or multimodal input. The self-supervised learning models do not require any labels during training.

3.2. Data Preprocessing

We provide a conceptual description of the preprocessing operations performed to create the multimodal datasets for all tasks. We note that all procedures described in this section are implemented based on recent work (Hayat et al., 2022; Harutyunyan et al., 2019) and are made readily available in our open-access repository.

3.2.1. Data Source

We used two large-scale publicly available datasets, MIMIC-IV (Johnson et al., 2023) and MIMIC-CXR (Johnson et al., 2019), to develop the proposed benchmark. MIMIC-IV consists of EHR data collected from patients admitted at the Beth Israel Deaconess Medical Center to the ICU between 2008 and 2019. In total, MIMIC-IV includes clinical data collected from 315,460 patients from 454,324 admissions and 76,943 ICU stays (Johnson et al., 2020). It also includes extensive information pertaining to the patient stay such as demographics, vital signs, laboratory test results, and procedure codes. MIMIC-CXR is composed of 377,110 CXR scans collected from 65,152 patients between 2011 and 2016, spanning over 227,835 studies. A waiver of informed consent was approved by the Institutional Review Board (IRB) to allow the sharing of this data. Additional information pertaining to the dataset can be found in the original work (Johnson et al., 2023, 2019).

Table 1: **Summary of existing medical multimodal benchmarks.** We briefly highlight the characteristics of existing benchmarks, including the medical scope and number of included tasks. Note that SSL is Self-Supervised Learning, IMG is Imaging, and QA is Question Answering.

Benchmark	Data Source	Scope	EHR	IMG	# Tasks	SSL
EHRXQA (Bae et al., 2024)†	MIMIC-IV (Johnson et al., 2020), MIMIC-CXR (Johnson et al., 2019), ImaGenome (Wu et al., 2021)	QA	✓	✓	1	×
ECG-QA (Oh et al., 2024)†	PTB-XL ECG (Wagner et al., 2020)	QA	✓	×	1	✓
EHRSHOT (Wornow et al., 2024)†	STARR (Callahan et al., 2023)	In-hospital tasks	✓	×	15	✓
MC-BEC (Chen et al., 2024a)†	Stanford MC-MED	Emergency Care	✓	×	3	×
RadFusion (Zhou et al., 2021)	STARR (Callahan et al., 2023)	Pulmonary Embolism	✓	✓	1	×
INSPECT (Huang et al., 2023)	STARR (Callahan et al., 2023)	Pulmonary Embolism	✓	✓	4	×
MIMIC-Extract (Harutyunyan et al., 2019)	MIMIC-III (Johnson et al., 2016)	Acute care	✓	×	4	×
EHR-TS-PT (McDermott et al., 2021)	MIMIC-III (Johnson et al., 2016), eICU (Pollard et al., 2018)	Acute care	✓	×	10	✓
MedFuse (Hayat et al., 2022)	MIMIC-IV (Johnson et al., 2020), MIMIC-CXR (Johnson et al., 2019)	Acute care	✓	✓	2	×
`MedMod` (Ours)	MIMIC-IV (Johnson et al., 2020), MIMIC-CXR (Johnson et al., 2019)	Acute care	✓	✓	5	✓

† Studies include other modalities, such as text and/or time-series data (e.g., ECG or PPG) that are not considered in our work.

3.2.2. EHR Feature Extraction

We extracted a diverse set of clinically relevant variables from the MIMIC-IV dataset that are critical for the prediction tasks in our benchmark. We utilized a consistent set of 17 clinical variables, comprising both categorical and continuous data (Harutyunyan et al., 2019; Hayat et al., 2022). The extracted variables include vital signs, laboratory measurements, and clinical assessment scores. Our selection was based on the proven predictive power of the selected variables in previous studies as well as the utilization frequency in clinical practice. We provide a description of these data items, including their source table and impute value in Appendix A. The following are the EHR preprocessing steps conducted:

- **Time-series Data Generation:** The EHR features extracted are collected at irregular intervals during the patient's ICU stay. Hence, to obtain a temporal representation of the data, we sample all extracted features at regular intervals, specifically every two hours. This step will result in an array of data such that $x_{ehr} \in \mathbb{R}^{d \times t}$, where d is the number of EHR features ($d = 17$) and t is the bi-hourly time step.
- **Handling Missing Data:** As a result of the bi-hourly data sampling, there will be a significant amount of missing data in each EHR timeseries. To overcome this issue, we employ two imputation strategies including (i) imputation with the most recent measurement and (ii) imputation with the normal value. The normal values used for imputation are summarized in Appendix A.
- **Data Discretization & Normalization:** As the initial set of clinical variables is composed of both categorical and continuous variables, we perform discretization to one-hot encode all categorical variables. This results in a data matrix consisting of 76 features encapsulating both categorical and continuous features, such that $x_{ehr} \in \mathbb{R}^{d \times t}$, where $d = 76$ and t is the time step. We also standardize the continuous variables to ensure consistency across different scales. Each task uses a task-specific normalizer where the statistics (mean and standard deviations) are computed based on the dataset.

3.2.3. CXR Filtering

We follow a series of preprocessing steps to filter the CXR images and verify that the selected images are clinically relevant and in accordance with the EHR data. Firstly, we only consider CXR scans that are gathered during the patient's current ICU stay, ensuring that the images reflect the patient's condition. Then, we only consider frontal scans with a view position of Anterior-Posterior (AP), as this is the standard perspective used for clinical assessment in critical care settings. Given that patients may have multiple CXR scans during their ICU stay, the utilization

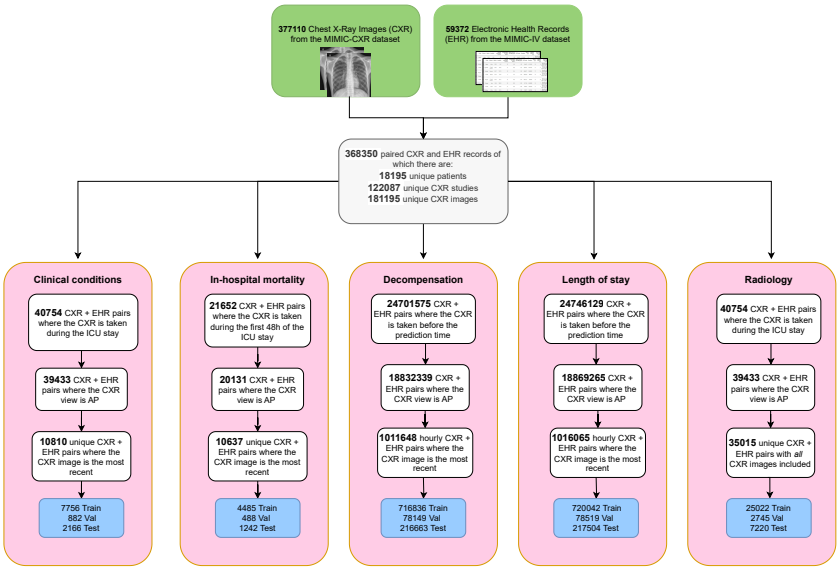

Figure 2: **Data Filtering Pipeline.** Overview of the data filtering process starting from the initial MIMIC-CXR and MIMIC-IV datasets to the creation of task-specific datasets for each of the five benchmark task. The initial pairing is common across all tasks and involves pairing CXR images and EHR records based on the ICU stay identification number (`stay_id`). The following filtering steps are task-specific, outlining the steps to generate the CXR-EHR pairs in the final task datasets.

of these images depends on the task, with two different filtering strategies employed. In the case of the radiological findings classification task, we consider all CXR scans gathered within an ICU stay, while for all other tasks we consider only the most recent CXR scan.

3.2.4. Multimodal Pairing

To construct our multimodal dataset from MIMIC-IV and MIMIC-CXR, we included paired samples only, where both the CXR and EHR modalities must be present in a given ICU stay. We used the patient identifier to randomly split our dataset into 70%, 10%, and 20% for the training, validation, and test sets, respectively. The resulting task label distributions is provided in Appendix B. Figure 2 presents a detailed description of the data filtering pipeline used to link the EHR and CXR data and create the multimodal dataset for each task.

3.3. Baselines

3.3.1. Supervised Learning

In the supervised learning setting, we evaluate all tasks using vanilla fusion techniques involving early, joint and late fusion. We also consider sophisticated fusion frameworks present in the literature, namely MedFuse, DAFT, and MMTM. For all models, we use an LSTM (Hochreiter and Schmidhuber, 1997) as an encoder for the EHR data, denoted as f_{ehr}, and a ResNet-34 (He et al., 2016) as an encoder for CXR data, denoted as f_{cxr}. Specifically, we pretrain f_{cxr} using the radiology labels from MIMIC-CXR, and pretrain f_{ehr} on the EHR labels for the respective task. Figure 1 provides a visual illustration of all supervised learning models. For the length of stay task, we use the Cross Entropy loss. For all other tasks, we use Binary Cross Entropy. We briefly summarize the supervised baselines.

Early fusion is the simplest form of fusion where modalities are fused at the input level and processed via a single encoder that models both modalities simultaneously (Huang et al., 2020). It pretrains unimodal models f_{ehr} and f_{cxr} separately, then fuses

the extracted features v_{ehr} and v_{cxr} as $v_{fusion} = [v_{ehr}; v_{cxr}]$ to learn from the combined data, with a classification layer g that computes the final output.

Joint Fusion is another type of fusion where representations learned from multiple modality-specific encoders are combined during training by means of a fusion layer (Huang et al., 2020). It extracts features using f_{ehr} for EHR and f_{cxr} for CXR, then concatenates v_{ehr} and v_{cxr} early in the network as $v_{fusion} = [v_{ehr}; v_{cxr}]$ to allow for joint learning without pretraining, with a classification layer g to produce the prediction.

Late Fusion operates on the output level by aggregating the predictions obtained from multiple modality-specific classifiers by averaging or majority voting (Huang et al., 2020). In our work, we average the predictions of both modalities.

MedFuse (Hayat et al., 2022) is a recent multimodal framework proposed for processing CXR scans and multivariate clinical time-series data. MedFuse replaces traditional fusion mechanisms with an LSTM-based fusion module that processes the joint representations of the input modalities.

Dynamic Affine Feature Map Transform (DAFT) (Pölsterl et al., 2021) is a multimodal fusion framework developed for processing medical images and tabular medical data. It operates on the dynamic conditioning (including feature map shifting and scaling) of the representations learned from medical images on a patient's tabular clinical information. It uses dynamic affine transformations on feature maps from modality-specific encoders. These transformations are conditioned on global context vectors derived from the input data. The transformed features are then fused, and a classification layer g is applied to the final fused representation for prediction.

Multimodal Transfer Module (MMTM) (Joze et al., 2020) is a multimodal fusion framework that operates on the intermediate layers of convolutional neural network (CNNs). The MMTM module is characterized by its ability to be located at different levels of the feature hierarchy, which allows for slow modality fusion. It uses modality-specific encoders f_{ehr} and f_{cxr} to generate feature representations v_{ehr} and v_{cxr} and uses squeeze and excitation operations to recalibrate channel-wise features from both modalities combining them into a joint representation to modulate the original features.

3.3.2. Self-supervised Learning

We apply three self-supervised learning methods (illustrated in Figure 1) to assess their ability in learning multimodal representations that are agnostic to task-specific labels. After pretraining without labels, we evaluate the quality of the learned representations via linear evaluation and fine-tuning for the five tasks proposed in `MedMod`.

Variance-Invariance-Covariance Regularization (VICReg) (Bardes et al., 2021) is a pretraining framework developed mainly to eliminate dimensional collapse in joint embedding frameworks. VICReg introduces regularization terms that prevent collapse by maintaining feature variance at certain thresholds and keeping the correlation between features minimal, while maintaining similarity. Regularization in VICReg is applied separately to each branch which renders it suitable for multimodal representation learning. The loss is defined as $\mathcal{L} = \mathcal{S}(z_{ehr}, z_{cxr})$, where $\mathcal{S}$ is the VICReg loss.

Contrastive VIsual Representation Learning from Text (ConVIRT) (Zhang et al., 2022) is a contrastive vision-language pretraining framework developed mainly for medical images and radiology reports. ConVIRT introduced a bidirectional contrastive objective between the pretraining modalities that maximizes the similarity of the embeddings of an image-text pair. The loss function of ConVIRT is defined as $\mathcal{L} = \mathcal{S}(z_{ehr}, z_{cxr}) + \mathcal{S}(z_{cxr}, z_{ehr})$, where $\mathcal{S}$ is the infoNCE loss.

A Large-scale ImaGe and Noisy-text embedding (ALIGN) (Jia et al., 2021) is a contrastive vision-language representation learning framework developed on natural image-caption pairs. The objective of ALIGN is to scale up the pretraining data by generating noisy image-caption pretraining pairs, while maintaining quality representations. The loss of the ALIGN framework is defined as $\mathcal{L} = \mathcal{S}(z_{ehr}, z_{cxr})$, where $\mathcal{S}$ is the infoNCE loss.

It is worth noting that both ConVIRT (Zhang et al., 2022) and ALIGN (Jia et al., 2021) are Contrastive Language-Image Pretraining (CLIP) approaches which we adapt for EHR-CXR data. In ConVIRT, matching pairs of CXR and EHR from the same patient are considered as positive pairs, while in ALIGN, noisy pairs of CXR and EHR (which may or may not come from the same patient) are considered as positive pairs. Further implementation details are described in Appendix C.

4. Results

We introduce a set of five clinical prediction tasks and perform evaluations of six supervised learning models and three self-supervised learning models, with a unified evaluation scheme followed for each task.

4.1. Overview of Tasks

We define five medical prediction tasks for `MedMod`, building upon prior benchmarks that only introduced unimodal tasks and two multimodal tasks (Harutyunyan et al., 2019; Hayat et al., 2022). Figure 1 provides a visual illustration of the proposed tasks and Table 2 provides a summary of their characteristics.

1. **In-hospital mortality prediction** is a binary classification task that involves predicting in-hospital mortality by the end of the first 48 hours of a patient's ICU stay. We consider Area Under the Receiver Operating Characteristic Curve (AUROC) and Area Under the Precision-Recall Curve (AUPRC) for evaluation of this task. Clinically, this task helps in identifying high-risk patients early, enabling interventions which could potentially lower mortality rates.

2. **Clinical conditions classification** is a multi-label classification task that aims at predicting the presence of any of 25 chronic, mixed, and acute care conditions, which are assigned to a patient at the end of an ICU stay. We use AUROC and AUPRC to evaluate model performance. The outcome of this task guides clinical decision-making and care.

3. **Decompensation prediction** is a binary prediction task that defines decompensation as mortality within the next 24 hours, computed at each hour of an ICU stay. The main metrics used for evaluating this task are AUROC and AUPRC. The aim is to replace early warning scores used in hospitals (Harutyunyan et al., 2019), and identify patients with deteriorating conditions.

4. **Length of stay prediction** is a multi-class classification task that entails predicting the patient's remaining time in the ICU at each hour of an ICU stay. The task involves the classification of the predicted length of stay values by sorting them into 10 buckets (Harutyunyan et al., 2019). To evaluate model performance on this task, we use Cohen's linear weighted kappa score (KAPPA) and Median Absolute Deviation (MAD). Although slightly different to the previous tasks, length of stay is essential in enabling better hospital management and better use of resources (Chen et al., 2024a), and has the potential to improve overall patient care.

5. **Radiological findings classification** is a multi-label classification task that involves predicting a set of 14 chest observations extracted from the available radiology reports of a CXR scan gathered during an ICU stay. The prediction is made at the end of the stay, with AUROC and AUPRC used to evaluate performance. With CXRs being one of the most widely utilized medical imaging exams, improving predictive performance for chest disease classification is a highly relevant clinical task.

Our selection of these tasks is driven by the goal of exploring how multimodal data can enhance predictive performance over a range of clinical scenarios. Each task was chosen with the consideration of the potential contribution of CXR imaging to the predictive model. For example, in relation to the length of stay task which might be an EHR oriented task, abnormalities present in a CXR scan may indicate complications that could extend the patient's ICU stay, making the task relevant to both modalities.

Importantly, we note that integrating EHR data with medical images has demonstrated performance improvements in multiple studies for several tasks such as length of stay (Chen et al., 2024b; Wang et al., 2024) and in-hospital mortality (Hayat et al., 2022; Wang et al., 2024; Khader et al., 2023). Therefore, all tasks introduced as part of `MedMod` are routine clinical prediction tasks which are essential for improving patient outcomes, and are widely researched within both the medical and machine learning communities.

4.2. Supervised Learning Results

Table 3 summarizes the results of the supervised baselines for the five tasks presented in `MedMod`. To evaluate the performance gain achieved through the multimodal fusion models, we include unimodal (EHR and CXR) results as well as multimodal results using different fusion strategies. We report AUROC and AUPRC for all tasks aside from length of stay, which is reported using KAPPA and MAD.

The multimodal models consistently outperform the unimodal models in all benchmark tasks. For

Table 2: **Overview of MedMod Benchmark Tasks.** We briefly summarize the main characteristics of the MedMod benchmark involving labels information, time horizon, training splits, and metrics.

Task	Label Type	# Labels	Time Horizon	Train/Val/Test	Evaluation
In-hospital mortality	Binary label	2	First 48 hours	4485/488/1242	AUROC, AUPRC
Clinical conditions	Multi-label	25	End of stay	7756/882/2166	AUROC, AUPRC
Decompensation	Binary label	2	Hourly	716836/78149/216663	AUROC, AUPRC
Length of stay	Multi-class	10	Hourly	720042/78519/217504	KAPPA, MAD
Radiology	Multi-label	14	End of stay	25022/2745/7220	AUROC, AUPRC

Table 3: **Performance results for each task comparing the unimodal and multimodal models in the supervised setting.** We report the (average ± std) KAPPA and MAD for the length of stay task and the (average ± std) AUROC and AUPRC for the remaining four tasks. Results for the best modality and best model for each task are in bold.

	In-hospital mortality		Clinical conditions		Decompensation		Length of stay		Radiology	
Method	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**KAPPA**	**MAD**	**AUROC**	**AUPRC**
Unimodal										
EHR (LSTM)	**0.829 ± 0.009**	**0.502 ± 0.028**	**0.720 ± 0.004**	**0.409 ± 0.006**	**0.862 ± 0.012**	**0.247 ± 0.009**	**0.380 ± 0.010**	**143.6 ± 2.3**	-	-
CXR (ResNet-34)	0.679 ± 0.007	0.246 ± 0.020	0.673 ± 0.006	0.360 ± 0.009	0.694 ± 0.007	0.037 ± 0.003	-	-	**0.705 ± 0.008**	**0.323 ± 0.006**
Multimodal (Pretrained)										
Early fusion	**0.842 ± 0.004**	**0.515 ± 0.020**	0.742 ± 0.008	0.431 ± 0.014	0.857 ± 0.011	0.154 ± 0.007	0.371 ± 0.008	455.1 ± 4.8	0.728 ± 0.012	0.338 ± 0.008
Late fusion	0.833 ± 0.009	0.472 ± 0.022	0.743 ± 0.010	0.427 ± 0.012	**0.868 ± 0.010**	**0.261 ± 0.014**	0.339 ± 0.004	140.6 ± 4.2	**0.732 ± 0.013**	**0.328 ± 0.009**
MedFuse	0.819 ± 0.007	0.482 ± 0.028	**0.744 ± 0.003**	**0.440 ± 0.010**	0.822 ± 0.014	0.178 ± 0.009	0.307 ± 0.009	140.9 ± 3.9	0.720 ± 0.011	0.334 ± 0.007
Multimodal (Random Initialization)										
Joint fusion	0.830 ± 0.008	0.499 ± 0.028	0.741 ± 0.002	0.433 ± 0.003	0.864 ± 0.013	0.245 ± 0.008	0.238 ± 0.007	150.2 ± 4.7	0.646 ± 0.013	0.290 ± 0.009
MMTM	0.783 ± 0.013	0.363 ± 0.032	0.721 ± 0.006	0.399 ± 0.019	0.844 ± 0.015	0.108 ± 0.005	0.261 ± 0.006	146.6 ± 4.5	0.653 ± 0.011	0.282 ± 0.010
DAFT	0.826 ± 0.008	0.494 ± 0.030	0.722 ± 0.004	0.414 ± 0.004	0.756 ± 0.015	0.070 ± 0.006	**0.417 ± 0.007**	**174.2 ± 4.6**	0.658 ± 0.012	0.294 ± 0.008

instance, early fusion shows the best performance for the in-hospital mortality task (0.842 ± 0.004), outperforming the unimodal EHR model (0.829 ± 0.009). Similarly, MedFuse (0.744 ± 0.003), a more sophisticated fusion technique, shows a noticable improvement in the clinical conditions task, compared to the unimodal EHR model (0.720 ± 0.004). This indicates that combining data from multiple modalities enhances the predictive power of the model and improves overall performance. Moreover, the pretrained multimodal performed better for four out of the five tasks when compared to the randomly initialized models.

However, while the multimodal models may have a significant advantage in tasks such as in-hospital mortality prediction and clinical conditions classification, they only marginally improved performance compared to the unimodal EHR model for the decompensation task, where late fusion (0.868 ± 0.010) performs comparably to the unimodal EHR model (0.862 ± 0.012). Therefore, further investigation into task-specific characteristics is needed to better explain the performance boost achieved by using multiple modalities.

4.3. Self-supervised Learning Results

We evaluate three self-supervised learning methods, ConVIRT (Zhang et al., 2022), VICReg (Bardes et al., 2021), and ALIGN (Jia et al., 2021) on their ability to learn representations that are agnostic to the downstream task. The quality of the learned representations is evaluated via linear probing and fine-tuning. Results for all three baselines on the multimodal dataset are reported in Table 4, while unimodal EHR and unimodal CXR results are in Table 5 and Table 6, respectively.

By comparing the results of linear probing to fine-tuning in Table 4, we observe only a marginal improvement from fine-tuning the multimodal model. For example, for both the in-hospital mortality and decompensation tasks, the best performing self-supervised method is ConVIRT using linear probing. This suggests that pre-trained generalized models can

Table 4: **Downstream performance results for linear probing and fine-tuning across the five benchmark tasks with multimodal data.** We report the (average ± std) KAPPA and MAD for the length of stay task and the (average ± std) AUROC and AUPRC for the remaining four tasks. We include the best supervised learning results obtained for each task in the multimodal setting, based on AUROC.

	In-hospital mortality		Clinical conditions		Decompensation		Length of stay		Radiology	
Method	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**KAPPA**	**MAD**	**AUROC**	**AUPRC**
Supervised learning	0.842 ± 0.004	0.515 ± 0.020	**0.744 ± 0.0003**	0.440 ± 0.010	0.868 ± 0.010	0.261 ± 0.014	**0.417 ± 0.007**	174.2 ± 4.6	**0.732 ± 0.013**	0.328 ± 0.009
Linear probing										
ConVIRT	**0.847 ± 0.002**	0.482 ± 0.016	0.701 ± 0.014	0.365 ± 0.015	**0.890 ± 0.011**	0.267 ± 0.012	0.373 ± 0.015	140.2 ± 4.3	0.661 ± 0.010	0.277 ± 0.016
VICReg	0.811 ± 0.008	0.458 ± 0.017	0.644 ± 0.018	0.322 ± 0.016	0.874 ± 0.012	0.233 ± 0.014	0.366 ± 0.013	142.0 ± 3.8	0.623 ± 0.011	0.251 ± 0.017
ALIGN	0.803 ± 0.014	0.464 ± 0.016	0.665 ± 0.017	0.359 ± 0.015	0.869 ± 0.013	0.258 ± 0.015	0.327 ± 0.014	143.5 ± 4.9	0.651 ± 0.009	0.274 ± 0.018
Fine-tuning										
ConVIRT	0.813 ± 0.010	0.449 ± 0.018	0.728 ± 0.016	0.422 ± 0.017	0.859 ± 0.010	0.088 ± 0.011	0.368 ± 0.014	101.3 ± 3.7	0.660 ± 0.012	0.287 ± 0.015
VICReg	0.834 ± 0.017	0.510 ± 0.017	0.693 ± 0.014	0.405 ± 0.016	0.832 ± 0.009	0.050 ± 0.009	0.379 ± 0.012	108.8 ± 4.3	0.637 ± 0.014	0.273 ± 0.016
ALIGN	0.811 ± 0.012	0.475 ± 0.019	0.711 ± 0.015	0.416 ± 0.018	0.807 ± 0.018	0.055 ± 0.012	0.374 ± 0.013	100.9 ± 4.4	0.628 ± 0.010	0.261 ± 0.015

Table 5: **Downstream performance results for linear probing and fine-tuning across the five benchmark tasks with EHR data.** We report the (average ± std) KAPPA and MAD for the length of stay task and the (average ± std) AUROC and AUPRC for the remaining four tasks. We include the supervised learning results obtained for each task in the unimodal EHR setting, based on AUROC.

	In-hospital mortality		Clinical conditions		Decompensation		Length of stay	
Method	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**KAPPA**	**MAD**
Supervised learning	0.829 ± 0.009	0.502 ± 0.028	0.720 ± 0.004	0.409 ± 0.006	0.862 ± 0.012	0.252 ± 0.005	0.380 ± 0.010	143.6 ± 2.3
Linear probing								
ConVIRT	0.789 ± 0.015	0.445 ± 0.012	0.664 ± 0.019	0.365 ± 0.013	0.873 ± 0.017	0.247 ± 0.020	0.360 ± 0.017	137.4 ± 6.2
VICReg	0.801 ± 0.009	0.437 ± 0.014	0.648 ± 0.016	0.321 ± 0.018	**0.884 ± 0.018**	0.264 ± 0.016	0.374 ± 0.015	140.1 ± 6.8
ALIGN	0.812 ± 0.020	0.453 ± 0.015	0.655 ± 0.021	0.335 ± 0.012	0.871 ± 0.014	0.253 ± 0.017	0.377 ± 0.018	138.7 ± 6.5
Fine-tuning								
ConVIRT	0.822 ± 0.019	0.453 ± 0.016	0.721 ± 0.017	0.425 ± 0.010	0.842 ± 0.013	0.090 ± 0.020	**0.389 ± 0.015**	103.2 ± 4.7
VICReg	0.827 ± 0.021	0.463 ± 0.014	**0.730 ± 0.020**	0.412 ± 0.009	0.853 ± 0.020	0.085 ± 0.018	0.384 ± 0.017	104.5 ± 5.1
ALIGN	**0.831 ± 0.006**	0.470 ± 0.017	0.715 ± 0.015	0.425 ± 0.013	0.845 ± 0.019	0.095 ± 0.016	0.377 ± 0.016	105.3 ± 5.4

perform well across most tasks without the need for additional and costly fine-tuning.

Additionally, by comparing the unimodal results in Tables 5 and 6, we observe that only using the EHR modality, compared to only using the CXR modality, resulted in consistently higher performance across all tasks except for the radiology task. This shows the relevance and inherent multimodality of the EHR data. Finally, the multimodal results presented in Table 4 compared with the unimodal results in Tables 5 and 6 highlight that the multimodal methods achieved improved results across all tasks. For example, ConVIRT trained on multimodal data is the best-performing model for the in-hospital mortality task (0.847 ± 0.002), a substantial improvement over the best model trained with EHR (0.831 ± 0.006) or CXR (0.687 ± 0.018) data only. This enhancement in performance indicates that multimodal data is especially important in obtaining generalized representations that are agnostic to the downstream task.

We note that we did not include results for the length of stay task based solely on CXR data, as our experiments showed that using CXR scans alone yielded very poor performance. This outcome aligns with clinical literature, which highlights the challenges of predicting length of stay without fine-grained temporal data and continuous physiological measurements, such as those present in EHRs (Wilk et al., 2020; Pungitore and Subbian, 2023).

5. Discussion

In this paper we presented `MedMod`, a comprehensive multimodal benchmark designed to facilitate advanc-

Table 6: **Downstream performance results for linear probing and fine-tuning across the five benchmark tasks with CXR data.** We report the (average ± std) KAPPA and MAD for the length of stay task and the (average ± std) AUROC and AUPRC for the remaining four tasks. We include the supervised learning results obtained for each task in the unimodal CXR setting, based on AUROC.

	In-hospital mortality		Clinical conditions		Decompensation		Radiology	
Method	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
Supervised learning	0.679 ± 0.007	0.246 ± 0.020	**0.673 ± 0.006**	0.360 ± 0.009	0.694 ± 0.011	0.037 ± 0.003	**0.705 ± 0.013**	0.323 ± 0.006
Linear probing								
ConVIRT	**0.687 ± 0.018**	0.275 ± 0.016	0.652 ± 0.021	0.326 ± 0.014	**0.723 ± 0.018**	0.037 ± 0.020	0.670 ± 0.015	0.275 ± 0.017
VICReg	0.641 ± 0.019	0.239 ± 0.017	0.589 ± 0.020	0.263 ± 0.016	0.613 ± 0.021	0.029 ± 0.018	0.604 ± 0.014	0.245 ± 0.019
ALIGN	0.675 ± 0.020	0.254 ± 0.015	0.621 ± 0.018	0.308 ± 0.017	0.524 ± 0.016	0.021 ± 0.014	0.610 ± 0.018	0.263 ± 0.016
Fine-tuning								
ConVIRT	0.620 ± 0.015	0.205 ± 0.018	0.660 ± 0.017	0.266 ± 0.014	0.534 ± 0.019	0.023 ± 0.016	0.638 ± 0.016	0.259 ± 0.018
VICReg	0.590 ± 0.020	0.212 ± 0.015	0.645 ± 0.016	0.222 ± 0.017	0.552 ± 0.018	0.019 ± 0.020	0.615 ± 0.017	0.261 ± 0.017
ALIGN	0.603 ± 0.017	0.198 ± 0.016	0.550 ± 0.020	0.259 ± 0.014	0.560 ± 0.020	0.017 ± 0.015	0.595 ± 0.018	0.233 ± 0.016

ing medical prediction tasks using EHR data and CXR scans. MedMod comprises five medical prediction tasks along with baseline results to provide a foundation for future research. MedMod addresses the critical need for multimodal benchmarks that reflect the complexities present in real-world clinical environments.

To the best of our knowledge, our benchmark serves as the first multimodal benchmark using EHR and CXR data and including both supervised and self-supervised baseline results. While there has been increasing attention to machine learning methods for medical applications, there has been limited work focusing on generating multimodal datasets (specifically for EHR and CXR) and introducing benchmark tasks that they can be evaluated on. As we make our code publicly-accessible, we aim to provide a valuable resource to be used by the research community.

Despite the comprehensiveness of our benchmark, it still possesses some limitations. First, our benchmark considers a single source of data represented by the MIMIC patient cohort, which might hinder the generalizability of the models developed and evaluated on the presented tasks. Additionally, our benchmark's reliance on the CXR and EHR modalities may lead to temporal misalignment, as EHR data is collected continuously while CXR data is collected discretely, making it challenging to ensure that both modalities are representative of the patient status at prediction time. Furthermore, some of the benchmark tasks such as decompensation and length of stay require massive computational resources which might limit their usage in the absence of sufficient computational resources. Lastly, we evaluated our benchmark on a relatively small number of baselines which may not cover all state-of-the-art models.

As a future research direction, we plan to enhance the comprehensiveness of MedMod by introducing additional clinically relevant tasks such as predicting ICU readmission (Barbieri et al., 2020), chronic obstructive pulmonary disease (COPD) (Wang et al., 2022) or assessing pulmonary edema severity (Horng et al., 2021). These tasks have mostly been evaluated using unimodal data, thus developing MedMod to produce comprehensive multimodal results will further support future research. In addition, radiology reports hold valuable context that describes the findings present in CXR scans. They have been used as an additional source of data in self-supervised pretraining (Tiu et al., 2022) as well as in disease detection tasks (Chauhan et al., 2020). Hence, we aim to expand the dimensionality of the modalities included in MedMod by incorporating radiology reports associated with CXR scans. Additionally, we would like to incorporate transformer-based architectures in MedMod, to help address challenges like temporal misalignment between EHR and CXR. Moreover, given that MedMod currently relies on fully paired data, we encourage future research to develop methods capable of handling missing modalities, a common scenario in clinical settings. These are promising future research directions that will enable more comprehensive comparisons, better reflect real-world scenarios, and contribute to the advancement of the field.

Acknowledgments

This work was supported by ASPIRE, the technology program management pillar of Abu Dhabi's Advanced Technology Research Council (ATRC), via the ASPIRE Precision Medicine Research Institute Abu Dhabi (ASPIREPMRIAD) award grant number VRI-20-10, and the NYUAD Center for Artificial Intelligence and Robotics, funded by Tamkeen under the NYUAD Research Institute Award CG010. The research was carried out on the High Performance Computing resources at New York University Abu Dhabi.

References

Saeed Amal, Lida Safarnejad, Jesutofunmi A Omiye, Ilies Ghanzouri, John Hanson Cabot, and Elsie Gyang Ross. Use of multi-modal data and machine learning to improve cardiovascular disease care. *Frontiers in cardiovascular medicine*, 9: 840262, 2022.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, et al. Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images. *Advances in Neural Information Processing Systems*, 36, 2024.

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.

Sebastiano Barbieri, James Kemp, Oscar Perez-Concha, Sradha Kotwal, Martin Gallagher, Angus Ritchie, and Louisa Jorm. Benchmarking deep learning architectures for predicting readmission to the icu and describing patients-at-risk. *Scientific reports*, 10(1):1111, 2020.

Batuhan Bardak and Mehmet Tan. Improving clinical outcome predictions using convolution over medical entities with multimodal learning. *Artificial Intelligence in Medicine*, 117:102112, 2021.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020.

Alison Callahan, Euan Ashley, Somalee Datta, Priyamvada Desai, Todd A Ferris, Jason A Fries, Michael Halaas, Curtis P Langlotz, Sean Mackey, José D Posada, et al. The stanford medicine data science ecosystem for clinical and translational research. *JAMIA open*, 6(3):ooad054, 2023.

Geeticka Chauhan, Ruizhi Liao, William Wells, Jacob Andreas, Xin Wang, Seth Berkowitz, Steven Horng, Peter Szolovits, and Polina Golland. Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*, pages 529–539. Springer, 2020.

Emma Chen, Aman Kansal, Julie Chen, Boyang Tom Jin, Julia Reisler, David E Kim, and Pranav Rajpurkar. Multimodal clinical benchmark for emergency care (mc-bec): A comprehensive benchmark for evaluating foundation models in emergency medicine. *Advances in Neural Information Processing Systems*, 36, 2024a.

Junde Chen, Yuxin Wen, Michael Pokojovy, Tzu-Liang Bill Tseng, Peter McCaffrey, Alexander Vo, Eric Walser, and Scott Moen. Multi-modal learning for inpatient length of stay prediction. *Computers in Biology and Medicine*, 171:108121, 2024b.

Yuanhong Chen, Fengbei Liu, Hu Wang, Chong Wang, Yuyuan Liu, Yu Tian, and Gustavo Carneiro. Bomd: bag of multi-label descriptors for noisy chest x-ray classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21284–21295, 2023.

Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle,

et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26:1045–1057, 2013.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

Peter D Dueben, Martin G Schultz, Matthew Chantry, David John Gagne, David Matthew Hall, and Amy McGovern. Challenges and benchmark datasets for machine learning in the atmospheric sciences: Definition, status, and outlook. *Artificial Intelligence for the Earth Systems*, 1(3):e210002, 2022.

Shaker El-Sappagh, Jose M Alonso, SM Riazul Islam, Ahmad M Sultan, and Kyung Sup Kwak. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for alzheimer's disease. *Scientific reports*, 11(1): 2660, 2021.

Hossein Estiri, Zachary H Strasser, Jeffy G Klann, Pourandokht Naseri, Kavishwar B Wagholikar, and Shawn N Murphy. Predicting covid-19 mortality with electronic medical records. *NPJ digital medicine*, 4(1):15, 2021.

Junyi Gao, Yinghao Zhu, Wenqing Wang, Zixiang Wang, Guiying Dong, Wen Tang, Hao Wang, Yasha Wang, Ewen M Harrison, and Liantao Ma. A comprehensive benchmark for covid-19 predictive modeling using electronic health records in intensive care. *Patterns*, 5(4), 2024.

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.

Nasir Hayat, Krzysztof J Geras, and Farah E Shamout. Medfuse: Multi-modal fusion with clinical time-series data and chest x-ray images. *arXiv preprint arXiv:2207.07027*, 2022.

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.

Lars Heiliger, Anjany Sekuboyina, Bjoern Menze, Jan Egger, and Jens Kleesiek. Beyond medical imaging-a review of multimodal deep learning in radiology. *Authorea Preprints*, 2023.

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

Gregory Holste, Song Wang, Ajay Jaiswal, Yuzhe Yang, Mingquan Lin, Yifan Peng, and Atlas Wang. Cxr-lt: Multi-label long-tailed classification on chest x-rays. *PhysioNet*, 5:19, 2023.

Steven Horng, Ruizhi Liao, Xin Wang, Sandeep Dalal, Polina Golland, and Seth J Berkowitz. Deep learning to quantify pulmonary edema in chest radiographs. *Radiology: Artificial Intelligence*, 3(2): e190228, 2021.

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, 2020.

Shih-Cheng Huang, Zepeng Huo, Ethan Steinberg, Chia-Chun Chiang, Matthew P Lungren, Curtis P Langlotz, Serena Yeung, Nigam H Shah, and Jason A Fries. Inspect: A multimodal dataset for pulmonary embolism diagnosis and prognosis. *arXiv preprint arXiv:2311.10798*, 2023.

Bogdan Ionescu, Henning Müller, Mauricio Villegas, Alba García Seco de Herrera, Carsten Eickhoff, Vincent Andrearczyk, Yashin Dicente Cid, Vitali Liauchuk, Vassili Kovalev, Sadid A Hasan, et al. Overview of imageclef 2018: Challenges, datasets and evaluation. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings 9*, pages 309–334. Springer, 2018.

Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in Neural Information Processing Systems*, 35:36722–36732, 2022.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv. *version 0.4). PhysioNet. https://doi.org/10.13026/a3wn-hq05*, 2020.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3 (1):1–9, 2016.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–8, 2019.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1): 1, 2023.

Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L Iuzzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13289–13299, 2020.

Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018.

Firas Khader, Jakob Nikolas Kather, Gustav Müller-Franzes, Tianci Wang, Tianyu Han, Soroosh Tayebi Arasteh, Karim Hamesch, Keno Bressem, Christoph Haarburger, Johannes Stegmaier, et al. Medical transformer for multimodal survival prediction in intensive care: integration of imaging and non-imaging data. *Scientific Reports*, 13(1): 10666, 2023.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020.

Adrienne Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5(1):171, 2022.

Felix Krones, Umar Marikkar, Guy Parsons, Adam Szmul, and Adam Mahdi. Review of multimodal machine learning approaches in healthcare. *arXiv preprint arXiv:2402.02460*, 2024.

Thomas J Littlejohns, Jo Holliday, Lorna M Gibson, Steve Garratt, Niels Oesingmann, Fidel Alfaro-Almagro, Jimmy D Bell, Chris Boultwood, Rory Collins, Megan C Conroy, et al. The uk biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nature communications*, 11(1):2624, 2020.

Matthew McDermott, Bret Nestor, Evan Kim, Wancong Zhang, Anna Goldenberg, Peter Szolovits, and Marzyeh Ghassemi. A comprehensive ehr time-series pre-training benchmark. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 257–278, 2021.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

Jungwoo Oh, Gyubok Lee, Seongsu Bae, Joonmyoung Kwon, and Edward Choi. Ecg-qa: A comprehensive question answering dataset combined with electrocardiogram. *Advances in Neural Information Processing Systems*, 36, 2024.

João Luiz Junho Pereira, Kate Smith-Miles, Mario Andrés Muñoz, and Ana Carolina Lorena. Optimal selection of benchmarking datasets for unbiased machine learning algorithm evaluation. *Data Mining and Knowledge Discovery*, 38(2):461–500, 2024.

Nick A Phillips, Pranav Rajpurkar, Mark Sabini, Rayan Krishnan, Sharon Zhou, Anuj Pareek, Nguyet Minh Phu, Chris Wang, Mudit Jain, Nguyen Duong Du, et al. Chexphoto: 10,000+ photos and transformations of chest x-rays for benchmarking deep learning robustness. In *Machine Learning for Health*, pages 318–327. PMLR, 2020.

Rimma Pivovarov, Adler J Perotte, Edouard Grave, John Angiolillo, Chris H Wiggins, and Noémie Elhadad. Learning probabilistic phenotypes from heterogeneous ehr data. *Journal of biomedical informatics*, 58:156–165, 2015.

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

Sebastian Pölsterl, Tom Nuno Wolf, and Christian Wachinger. 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*, pages 688–698. Springer, 2021.

Hoifung Poon. Multimodal generative ai for precision health. *NEJM AI Sponsored*, 2023.

Sarah Pungitore and Vignesh Subbian. Assessment of prediction tasks and time window selection in temporal modeling of electronic health record data: a systematic review. *Journal of Healthcare Informatics Research*, 7(3):313–331, 2023.

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. Ai in health and medicine. *Nature medicine*, 28(1):31–38, 2022.

Benjamin A Satterfield, Ozan Dikilitas, and Iftikhar J Kullo. Leveraging the electronic health record to address the covid-19 pandemic. *Mayo Clinic Proceedings*, 96(6):1592–1608, 2021.

Thanveer Shaik, Xiaohui Tao, Lin Li, Haoran Xie, and Juan D Velásquez. A survey of multimodal information fusion for smart healthcare: Mapping the journey from data to wisdom. *Information Fusion*, page 102040, 2023.

Farah E Shamout, Yiqiu Shen, Nan Wu, Aakash Kaku, Jungkyu Park, Taro Makino, Stanislaw Jastrzkbski, Jan Witowski, Duo Wang, Ben Zhang, et al. An artificial intelligence system for predicting the deterioration of covid-19 patients in the emergency department. *NPJ digital medicine*, 4(1):80, 2021.

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.

Luis R Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M Wiberg, Michael L Li, Ignacio Fuentes, and Dimitris Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ digital medicine*, 5(1):149, 2022.

Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. *IEEE journal of biomedical and health informatics*, 25(5):1519–1528, 2020.

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1406, 2022.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):1–15, 2020.

Ryan Wang, Li-Ching Chen, Lama Moukheiber, Mira Moukheiber, Dana Moukheiber, Zach Zaiman, Sulaiman Moukheiber, Tess Litchman, Kenneth Seastedt, Hari Trivedi, et al. Early diagnosis of chronic obstructive pulmonary disease from chest x-rays using transfer learning and fusion strategies. *arXiv preprint arXiv:2211.06925*, 2022.

Yuanlong Wang, Changchang Yin, and Ping Zhang. Multimodal risk prediction with physiological signals, medical images and clinical notes. *Heliyon*, 10(5), 2024.

Marta Wilk, D William R Marsh, Sarah De Freitas, and John Prowle. Predicting length of stay in hospital using electronic records available at the time of admission. In *Digital Personalized Health and Medicine*, pages 377–381. IOS Press, 2020.

Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason Fries, and Nigam Shah. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. *Advances in Neural Information Processing Systems*, 36, 2024.

Joy Wu, Nkechinyere Agu, Ismini Lourentzou, Arjun Sharma, Joseph Paguio, Jasper Seth Yao, Edward Christopher Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. Chest imagenome dataset. *Physio Net*, 2021.

Feng Xie, Jun Zhou, Jin Wee Lee, Mingrui Tan, Siqi Li, Logasan S/O Rajnthern, Marcel Lucas Chee, Bibhas Chakraborty, An-Kwok Ian Wong, Alon Dagan, et al. Benchmarking emergency department prediction models with machine learning and public electronic health records. *Scientific Data*, 9 (1):658, 2022.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022.

Shuai Zheng, Zhenfeng Zhu, Zhizhe Liu, Zhenyu Guo, Yang Liu, Yuchen Yang, and Yao Zhao. Multi-modal graph learning for disease prediction. *IEEE Transactions on Medical Imaging*, 41(9): 2207–2216, 2022.

Yuyin Zhou, Shih-Cheng Huang, Jason Alan Fries, Alaa Youssef, Timothy J Amrhein, Marcello Chang, Imon Banerjee, Daniel Rubin, Lei Xing, Nigam Shah, et al. Radfusion: Benchmarking performance and fairness for multimodal pulmonary embolism detection from ct and ehr. *arXiv preprint arXiv:2111.11665*, 2021.

Appendix A. EHR Data Items

The extracted EHR data items are shown in Table 7. All variables are extracted from the *chartevents* table, except for Glucose and pH laboratory measurements which were extracted from the *labevents* tables. All features are collected during the patient ICU stay.

Table 7: **EHR data items**. A summary of the EHR data items utilized in `MedMod`, including their types, source tables, impute values, and a brief description of each variable.

#	Variable Name	Variable Description	Source Table	Impute Value
Categorical Variables				
1	Capillary Refill Rate	Indicator of circulatory system function	chartevents	0.0
2	Glasgow Coma Scale - Eye Opening	Assesses eye response to stimuli	chartevents	4 Spontaneously
3	Glasgow Coma Scale - Motor Response	Assesses motor response to stimuli	chartevents	6 Obeys Commands
4	Glasgow Coma Scale - Verbal Response	Assesses verbal response to stimuli	chartevents	5 Oriented
5	Glasgow Coma Scale - Total	Overall assessment of consciousness level	chartevents	15
Continuous Variables				
6	Diastolic Blood Pressure	Blood pressure during heart's relaxation phase	chartevents	59.0
7	Fraction of Inspired Oxygen	Oxygen concentration in inhaled air	chartevents	0.21
8	Glucose	Blood sugar level	labevents	128.0
9	Heart Rate	Number of heartbeats per minute	chartevents	86
10	Height	Patient's height	chartevents	170.0
11	Mean Blood Pressure	Average blood pressure during a single cardiac cycle	chartevents	77.0
12	Oxygen Saturation	Percentage of oxygen-saturated hemoglobin	chartevents	98.0
13	Respiratory Rate	Number of breaths per minute	chartevents	19
14	Systolic Blood Pressure	Blood pressure during heart's contraction phase	chartevents	118.0
15	Temperature	Body temperature	chartevents	36.6
16	Weight	Patient's weight	chartevents	81.0
17	pH	Acidity or alkalinity of the blood	labevents	7.4

Appendix B. Task Distribution

The label distributions for the benchmark tasks are shown in Figure 3. The figure shows the labels used for each of the tasks, including 25 phenotypes, 14 radiology classes, 10 length of stay buckets, and binary mortality labels (for both in-hospital mortality and decompensation). These distributions pertain to the test set of each task.

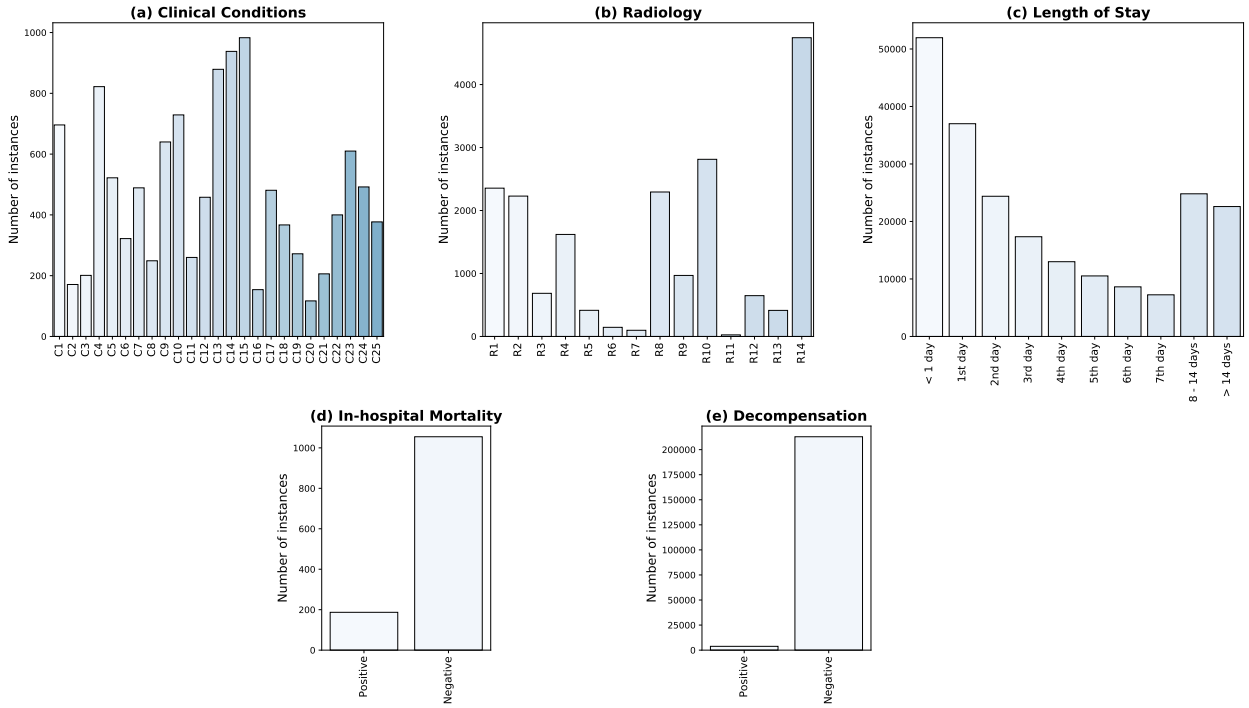

Figure 3: **Label Distribution for Benchmark Tasks. (a)** The distribution of the classes for the clinical conditions task is presented. The 25 clinical conditions labels displayed are: Acute and unspecified renal failure (C1), Acute cerebrovascular disease (C2), Acute myocardial infarction (C3), Cardiac dysrhythmias (C4), Chronic kidney disease (C5), Chronic obstructive pulmonary disease (C6), Complications of surgical procedures (C7), Conduction disorders (C8), Congestive heart failure (C9), Coronary atherosclerosis (C10), Diabetes mellitus with complications (C11), Diabetes mellitus without complication (C12), Disorders of lipid metabolism (C13), Essential hypertension (C14), Fluid and electrolyte disorders (C15), Gastrointestinal hemorrhage (C16), Hypertension with complications (C17), Other liver diseases (C18), Other lower respiratory disease (C19), Other upper respiratory disease (C20), Pleurisy; pneumothorax (C21), Pneumonia (except tuberculosis) (C22), Respiratory failure (C23), Septicemia (C24), Shock (C25). **(b)** The label distribution of the 14 chest observations for the radiology task is shown. The abbreviated labels of the observations are: Atelectasis (R1), Cardiomegaly (R2), Consolidation (R3), Edema (R4), Enlarged cardiomediastinum (R5), Fracture (R6), Lung lesion (R7), Lung opacity (R8), No finding (R9), Pleural effusion (R10), Pleural other (R11), Pneumonia (R12), Pneumothorax (R13), Support devices (R14). **(c)** The 10 length of stay buckets, showing the remaining duration of stay in the ICU. **(d)** The distribution of the binary labels for the in-hospital mortality task is shown, where the 'positive' label indicates mortality and the 'negative' label indicates the patient is alive. **(e)** The distribution of the hourly mortality labels for the decompensation task are displayed, with 'positive' indicating mortality and 'negative' indicating the patient is alive.

Appendix C. Implementation Details

Unimodal supervised training settings. We conducted thorough sweeps for training the unimodal EHR (LSTM) and CXR (ResNet-34) models to select the learning rate, by sampling from a uniform distribution in the range $[10^{-6}, 10^{-2}]$. This involved 100 runs for mortality, 20 runs for clinical conditions classification, and 10 runs each for the length-of-stay and decompensation tasks. For mortality and clinical conditions, the number of epochs was fixed as 50, while for decompensation and length-of-stay, it was set as 10. For MedFuse, early, and joint fusion, we pretrained the CXR encoder with the radiology labels for 100 epochs as in previous work (Hayat et al., 2022), and trained on the full CXR dataset.

Multimodal supervised fusion settings. We ran 100 hyperparameter tuning runs for mortality (50 epochs), 20 for clinical conditions (50 epochs), 10 for decompensation (10 epochs), 10 for length of stay (10 epochs), and 5 for radiology (50 epochs), by sampling the learning rate from a uniform distribution in the range $[10^{-6}, 10^{-2}]$. The batch size was fixed at 16 for all runs. We selected the best model checkpoint based on the epoch with the highest AUROC on the validation set. We implemented early stopping if the validation AUROC did not improve for 10 epochs (excluding the decompensation and length of stay tasks).

Self-supervised pre-training settings. We conducted 10 hyperparameter tuning runs for all baselines via random search by sampling a learning rate from a uniform distribution in the range $[10^{-1}, 10^{-2}]$. We used a batch size of 256 across all experiments, set the maximum number of epochs to 300, and introduced early stopping if the validation loss did not improve for 30 epochs. To select the best model for the downstream tasks, we ran linear evaluation for each pre-training epoch and chose the best checkpoint based on the best AUROC score achieved on the validation set across epochs and models.

Self-supervised linear evaluation/fine-tuning settings. For the linear evaluation and fine-tuning set-up, we conducted 5 runs of hyperparameter tuning with learning rates sampled from a uniform distribution in the range of $[10^{-4}, 10^{-1}]$. We used the Adam optimizer and ran experiments for 300 epochs and batch size 256 with early stopping implemented if validation AUROC did not improve for 10 epochs. We reported the best evaluation results achieved on the test set.

Chunk-wise training. We consider the chunk-wise training approach for the decompensation and length of stay tasks following the work of (Harutyunyan et al., 2019). Chunk-wise training involves training the model on a variable subset of the data and reporting metrics at the end of every N chunks instead of every N epochs. We consider this training strategy for these two tasks due to their large dataset size that requires a longer training time and due to the model overfitting before iterating over the full dataset.

Experimental details. All experiments were conducted using NVIDIA A100/V100 GPUs provided through an internal cluster. Each experiment is conducted with five random seeds, and the results presented in the respective results tables reflect the averages of the five independent runs.

Appendix D. Computational Cost

To explore how the computational cost of the different fusion models compare, we compute the number of parameters and average inference time for various multimodal fusion frameworks and present the results in Table 8. The number of parameters correspond to the trainable parameters of each fusion method. The average inference time is computed per batch, with the batch size fixed to 16.

In regard to the model size, the number of trainable parameters seems to be comparable among the advanced fusion techniques (MedFuse, MMTM, DAFT) as well as joint fusion (which involves joint training of the two modalities without pre-trained encoders). Self-supervised fine-tuning also involves a similar number of parameters. A similar trend is observed in terms of inference time, which is within the same range for all the supervised models. However, fine-tuning has a notably longer inference time, while linear evaluation is over 500x faster compared to the supervised methods.

To further analyze whether there exists a trade-off between the computational cost (as approximated by model size) and performance, we also include AUROC results for the in-hospital mortality task in Table 8. We find that though self-supervised fine-tuning has the smallest number of parameters among the advanced fusion techniques, it outperforms all other methods in this task. Notably, Early Fusion has nearly 20 times fewer parameters than the other supervised techniques but achieves the strongest performance in the supervised setting. We also highlight that linear evaluation involves an extremely minimal number of trainable parameters and achieves a significantly shorter inference time while delivering a strong performance that is comparable to the other fusion methods. Thus, in terms of deployment, utilizing linear evaluation of self-supervised pre-trained appears to be a promising approach, balancing performance, inference speed and computational cost.

Table 8: **Computational cost and inference time.** A summary of the number of trainable parameters, average inference time per batch (with a fixed batch size of 16), and AUROC on the in-hospital mortality task for a set of supervised and self-supervised multimodal fusion models.

Model	**Parameters (millions)**	**Inference time per batch (seconds)**	**AUROC**
MedFuse	23.9	0.0316	$0.819_{\pm 0.007}$
MMTM	23.5	0.0328	$0.783_{\pm 0.013}$
DAFT	22.3	0.0298	$0.826_{\pm 0.008}$
Early Fusion	1.70	0.0318	$0.842_{\pm 0.004}$
Joint Fusion	23.9	0.0321	$0.830_{\pm 0.008}$
Late Fusion	N/A	0.0284	$0.833_{\pm 0.009}$
ConVIRT (Fine-tune)	21.4	0.2232	$0.847_{\pm 0.002}$
ConVIRT (Linear evaluation)	0.000641	0.00004875	$0.813_{\pm 0.010}$

Appendix E. Transformer-based Results

To enhance the flexibility of our benchmark, we incorporate an option in our codebase to switch encoders to alternative architectures, specifically the transformer architecture. We provide results of a subset of the presented multimodal fusion approaches using transformer-based encoders for both the CXR and EHR modalities in Table 9.

Table 9: **Performance results using a transformer backbone for the in-hospital mortality, clinical conditions, and radiology tasks.** The table shows results using transformer-based encoders for both the CXR and EHR modalities.

Model	In-hospital mortality		Clinical Conditions		Radiology	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Early Fusion	0.8064	0.4481	**0.7760**	**0.5812**	0.6035	0.2786
Joint Fusion	0.8001	**0.4695**	0.7601	0.5574	0.7332	0.3893
Late Fusion	**0.8092**	0.3979	0.7713	0.5603	**0.7453**	**0.4031**

Appendix F. Calibration Scores

We provide calibration scores (Expected Calibration Error, ECE) for the unimodal and multimodal fusion models on the in-hospital mortality task in Table 10. MMTM achieves the lowest ECE of 0.518, indicating the best calibration among all methods. MedFuse also demonstrates comparatively competitive calibration with an ECE of 0.666, outperforming the unimodal models (ECE: 0.762 for EHR; 0.699 for CXR). These results indicate the advantages of multimodal fusion in improving reliability.

Table 10: **Calibration scores.** Expected Calibration Error (ECE) of unimodal and multimodal fusion models for predictions made on the in-hospital mortality task.

Model	ECE
Uni-modal EHR	0.762
Uni-modal CXR	0.699
MedFuse	0.666
MMTM	0.518
Daft	0.732
Early Fusion	0.741
Joint Fusion	0.774
Late Fusion	0.519

Appendix G. Task-specific Multimodal Improvement

We evaluate the performance improvement achieved by the multimodal models compared to the best-performing unimodal baselines across the five benchmark tasks. As shown in Table 11, while the multimodal models consistently outperform the unimodal models across all tasks, the percentage improvement is varying according to task-specific characteristics.

Table 11: **Performance improvement as a result of multimodal modelling.** Performance gain reported in percentage, based on the best unimodal and multimodal model for each of the five benchmark tasks.

Task	Unimodal result	Multimodal result	Gain (%)
In-hospital Mortality	0.829	0.842	1.57
Clinical Conditions	0.720	0.744	3.33
Decompensation	0.862	0.868	0.69
Length of Stay	0.380	0.417	9.74
Radiology	0.705	0.732	3.83

Appendix H. Interpretability Analysis

Figure 4 presents Grad-CAM visualizations of the unimodal CXR baseline (ResNet-34) and a multimodal fusion technique (MedFuse) for predictions made on the radiological classifications findings task across the following labels: Pleural Effusion, Support Devices, and No Finding.

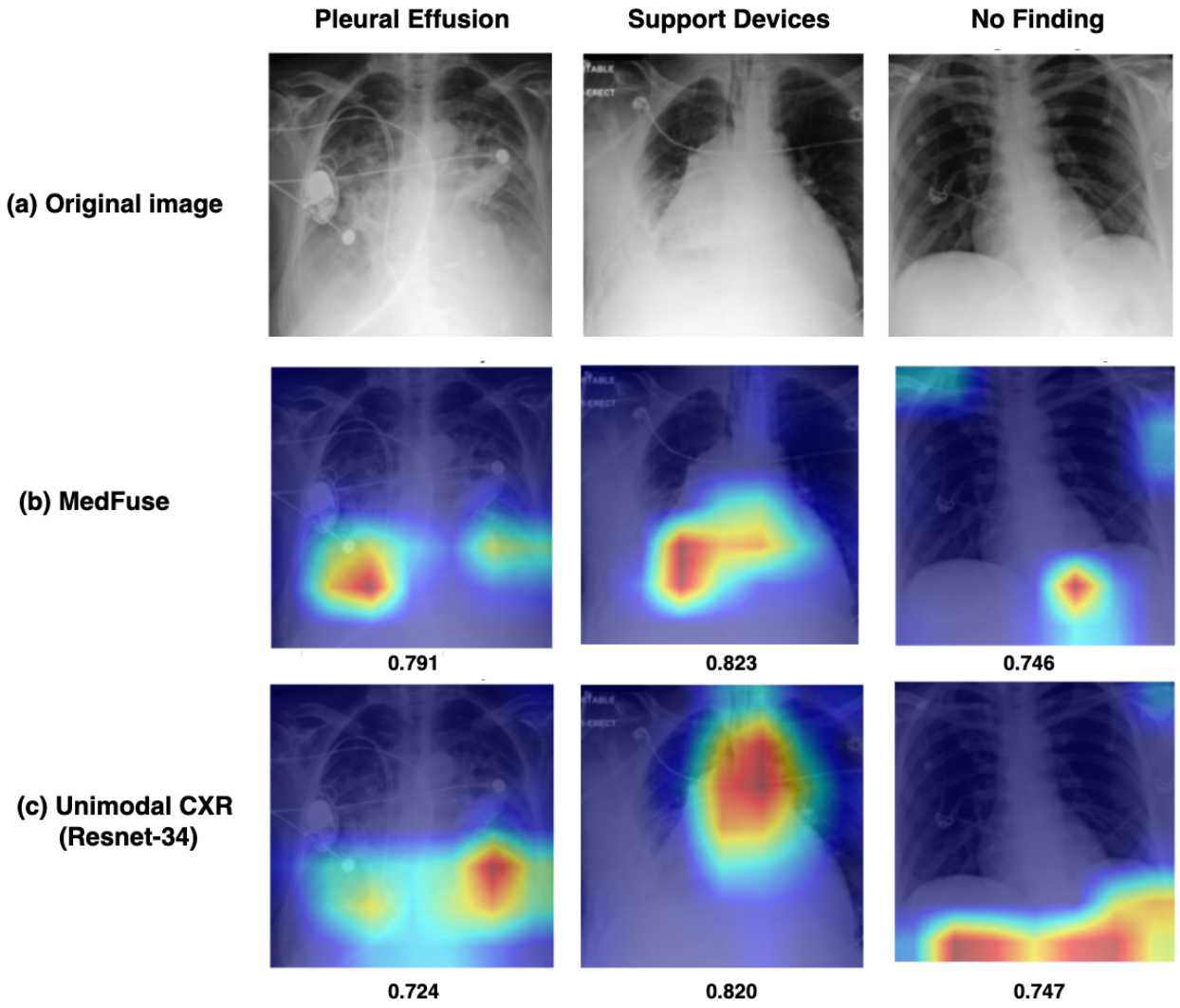

Figure 4: **Grad-CAM visualizations on the radiological findings classification task.** The generated Grad-CAM visualizations and respective predictions for the unimodal CXR and MedFuse baselines across the following labels: Pleural Effusion, Support Devices, and No Finding.