Equitable Electronic Health Record Prediction with FAME: Fairness-Aware Multimodal Embedding

Nikkie Hooman[1] NIKKIEH@SMU.EDU
Zhongjie Wu[1] ZHONGJIEW@SMU.EDU
Eric C. Larson[1,2] ECLARSON@SMU.EDU
Mehak Gupta[1] MEHAKG@SMU.EDU

[1] *Department of Computer Science, Southern Methodist University, Dallas, TX, USA*
[2] *Institute for Computational Biosciences, Southern Methodist University, Dallas, TX, USA*

Abstract

Electronic Health Record (EHR) data encompasses diverse modalities—text, images, and medical codes—that are vital for clinical decision-making. To process these complex data, multimodal AI (MAI) has emerged as a powerful approach for fusing such information. However, most existing MAI models optimize for better prediction performance, potentially reinforcing biases across patient subgroups. Although bias reduction techniques for multimodal models have been proposed, the individual strengths of each modality and their interplay in both reducing bias and optimizing performance remain underexplored. In this work, we introduce FAME (Fairness-Aware Multimodal Embeddings), a framework that explicitly weights each modality according to its fairness contribution. FAME optimizes both performance and fairness by incorporating a combined loss function. We leverage the Error Distribution Disparity Index (EDDI) to measure fairness across subgroups and propose a sign-agnostic aggregation method to balance fairness across subgroups, ensuring equitable model outcomes. We evaluate FAME with BEHRT and BioClinicalBERT, combining structured and unstructured EHR data, and demonstrate its effectiveness in performance and fairness compared to other baselines across multiple EHR prediction tasks.

1. Introduction

Healthcare decisions involve one or more physicians and health specialists that must interpret a myriad of data sources. Electronic health records (EHRs) offer a storage medium for this myriad of patient specific information across historical visits. Therefore, leveraging artificial intelligence (AI) in conjunction with EHRs holds transformative potential to improve healthcare. EHRs usually contains both structured (e.g., numerical, categorical) and unstructured data (e.g., text or image). Despite the routine use of EHR data to contextualize clinical history and inform medical decision-making, the majority of deep learning architectures are unimodal (Kline et al., 2022)—they only learn features from either structured or unstructured EHR data (Zhou et al., 2021). Multimodal AI (MAI) has emerged as a technique for analyses that can combine multiple types of data (e.g., text, images, medical codes) simultaneously, rather than relying solely on one modality, to generate outputs.

While multimodal AI (MAI) has gained traction in healthcare applications (Shaik et al., 2023; Cui et al., 2024), relatively little attention has been directed toward leveraging these

advancements to promote fairness. Although existing work increasingly evaluates MAI models for fairness, few studies explicitly explore how the integration and interaction of multiple modalities can be harnessed to achieve more equitable outcomes. Our work addresses this research gap by proposing fairness-aware fusion techniques. Our proposed techniques weigh each modality based on its prediction and fairness performance to build efficient and fairer multimodal AI models. We assess fairness across attributes such as ethnicity, age, and insurance type using EDDI (Error Distribution Disparity Index) (Wang et al., 2024) across all subgroups. We propose a sign-agnostic aggregation method to combine EDDI across all subgroups that can be employed not only in the loss function for our model, but also as part of a feedforward weighting scheme to ensure modalities that promote fairness are given priority in the model. To test our approach, we utilize two language encoders designed for healthcare data: BEHRT and BioClinicalBERT. BEHRT is a transformer-based model that has performed well in analyzing structured longitudinal EHR data and BioClinicalBERT is a specialized language model for unstructured clinical text. Combining these modalities ensures our model can process the myriad of input data from EHRs, while also providing a test case for our proposed EDDI weighting scheme. Our contributions are as follows:

1. We propose fairness-aware multimodal embeddings (FAME), a method for fusing multiple modalities in EHR data using weighted aggregation.

2. We introduce a method to derive fairness-aware weights using a sign-agnostic aggregation across EDDI values within a specified subgroup. Additionally, we implement a loss function that incorporates the aggregated EDDI to optimize the model while ensuring equality across subgroups.

3. We demonstrate our method's effectiveness through a series of experiments on three EHR prediction tasks, comparing it to other baselines and within an ablation of the model elements.

Generalizable Insights about Machine Learning in the Context of Healthcare

Existing multimodal AI models in healthcare primarily integrate multiple data types to enhance predictive performance, but they often do not explicitly account for how individual modalities contribute to fairness. While multimodal approaches can capture richer patient information, they typically treat different data sources as complementary inputs rather than considering their distinct roles in mitigating biases. Our work explores an alternative perspective by incorporating modality-specific weighting to promote fairness, demonstrating that structured and unstructured data can be leveraged not only for improving accuracy but also for reducing disparities across patient subgroups. This suggests that a more intentional approach to multimodal fusion—one that balances both predictive performance and fairness—may be beneficial in the development of equitable healthcare AI systems.

2. Related Work

In this section, we review existing bias mitigation methods, fusion methods, and fairness evaluation metrics in multimodal EHR models.

2.1. Bias Mitigation in Multimodal EHR models

The existence of biases and disparity across ethnicity, age, and other factors is a major concern in healthcare with the potential to be exacerbated by digital technologies and AI (Agarwal et al., 2023). Recent policy discussions have noted that with the growing impact of AI on business and society, it is even more critical to ensure that AI is fair, equitable, and treats all "populations" in an equivalent manner (Schwartz et al., 2022; Sharman, 2022; Mehrabi et al., 2021). This is especially important for healthcare, as health inequity can have life-altering consequences.

Recent studies have proposed various strategies to reduce the impact of patient demographics and socioeconomic factors on AI model outputs. Adversarial learning frameworks, for example, have been employed to mitigate bias in clinical prediction by learning representations that are invariant to sensitive attributes (Liu et al., 2022; Pfohl et al., 2021; Sivarajkumar et al., 2023). Similarly, contrastive learning techniques reduce contrastive loss between demographically based counterfactuals to promote performance parity across subgroups (Agarwal et al., 2024; Wang et al., 2024).

In parallel, several multimodal EHR models have begun to incorporate interpretability mechanisms to better understand the contributions and interactions of individual modalities. For instance, Lyu et al. (2023) proposed a multimodal transformer that jointly models clinical notes and structured data, emphasizing the distinct predictive strengths of each modality. Tsai et al. (2020) introduced multimodal routing to enhance both local and global interpretability, allowing examination of modality interactions during decision-making.

Despite these advances, most approaches apply debiasing techniques only after modality fusion, overlooking the unique fairness contributions of each modality. Our method explicitly evaluates and leverages the fairness of embeddings from individual modalities prior to fusion, enabling a more targeted and interpretable approach to multimodal bias mitigation.

2.2. Multimodal fusion methods

Several different strategies can be leveraged to fuse features from different modalities, including early fusion, late fusion, and joint fusion (Huang et al., 2020a,b; Zhou et al., 2021; Baltrušaitis et al., 2018). Early fusion combines features from separate modalities at the input level. Late fusion trains separate models for each modality and aggregate the predicted probability from all single-modality models as the final prediction. In joint fusion, intermediate representation (embeddings) from unimodal models (Vaswani et al., 2017; Devlin et al., 2019; Dosovitskiy et al., 2020; Huang et al., 2019; Raffel et al., 2020) are combined either through concatenation, addition, or MLP (multi-layer perceptron) fusion to obtain multimodal embeddings for each sample in the dataset.

Recent advances in deep learning technologies have led to the development of complex multimodal AI models in healthcare (Kline et al., 2022; Wang et al., 2024; Shaik et al., 2023; Cui et al., 2024; Rahman et al., 2020; Radford et al., 2021). Most of these existing techniques use multimodal fusion techniques like joint fusion or late fusion to combine embeddings and train models on performance-related metrics such as binary-cross entropy loss to produce high-performance models.

In our work, we focus on joint fusion and modify it to implement fairness-aware joint fusion. In existing joint fusion methods, embeddings from all modalities are either concate-

nated or averaged without weighting, which limits the utilization of the unique strengths of each modality. In our proposed method, FAME, we use fairness-metric information to weigh the unimodal embeddings before fusing them. This can help to control the contribution of each modality based on its observed fairness and leverage the joint fusion mechanisms not only to facilitate strong prediction but also to facilitate fair outcomes.

2.3. Fairness evaluation in Multimodal EHR models

Traditional fairness metrics like equalized odds and disparity index assess fairness for different subgroups in each demographic category (Hardt et al., 2016). According to these metrics, a model is considered fair if the error rate, like the true positive rate and false positive rate, across all subgroups is consistent. This measure ignores the relative size of subgroups and imbalanced outcomes in health data. From the wide range of metrics used in the community (Castelnovo et al., 2022), we use EDDI, specifically designed for datasets where subgroups and outcomes are imbalanced. Error Distribution Disparity Index (EDDI) (Hardt et al., 2016) measures the difference in error rates (proportion of incorrect prediction) between the privileged and unprivileged groups for a given prediction task. EDDI can be positive or negative depending on if subgroup error rate is higher or lower than the overall error rate, respectively. To obtain one EDDI value per category (e.g., race, insurance type) that ensures equality across all subgroups, we propose a sign-agnostic aggregation method to aggregate EDDI across each subgroup (e.g., White, Asian, Black) in a demographic category (e.g., race).

3. Methods

3.1. Problem Formulation

We define a multimodal dataset $D = (X_0, X_1, ...X_m..., X_M)_{m=1}^{M}$, where M is the number of modalities and X_m is feature data for all data points in the m^{th} modality. Our objective is to develop an effective and fair multimodal EHR model $f_{multi} : (X_0, X_1, ...X_m..., X_M)_{m=1}^{M} \rightarrow Y$, where $Y \in (0, 1)$ are the output binary labels. While training f_{multi}, we aim to effectively combine multiple modalities to improve prediction performance while also ensuring equality among all subgroups, S. Though our proposed methods can be applied to any number of modalites. In this work, we use $M = 3$ (i.e., three modalities) where X_d is demographic data, X_s is longitudinal clinical data (structured), X_n is longitudinal clinical notes (unstructured). We use a subset of demographic data as a set of sensitive attributes $S \subset X_d$. The subgroups, S, are selected using attributes identified as sensitive and, therefore, requiring fair outcomes across each group. The selection of these sensitive groupings is discussed in more detail in Section 4.3.

Our proposed approach uses the joint fusion technique to combine multiple modalities. Joint (or intermediate) fusion trains a decision-making model using extracted features from single-modality models while propagating the loss back to the unimodal feature-extracting models (Zhou et al., 2021). These extracted features are intermediate n-dimensional latent embeddings, z_m, that can be learned using any neural network. The input to each single-modality (or unimodal) model is only one type of data, X_m. These latent embeddings, z_m are then fused to predict the outcome.

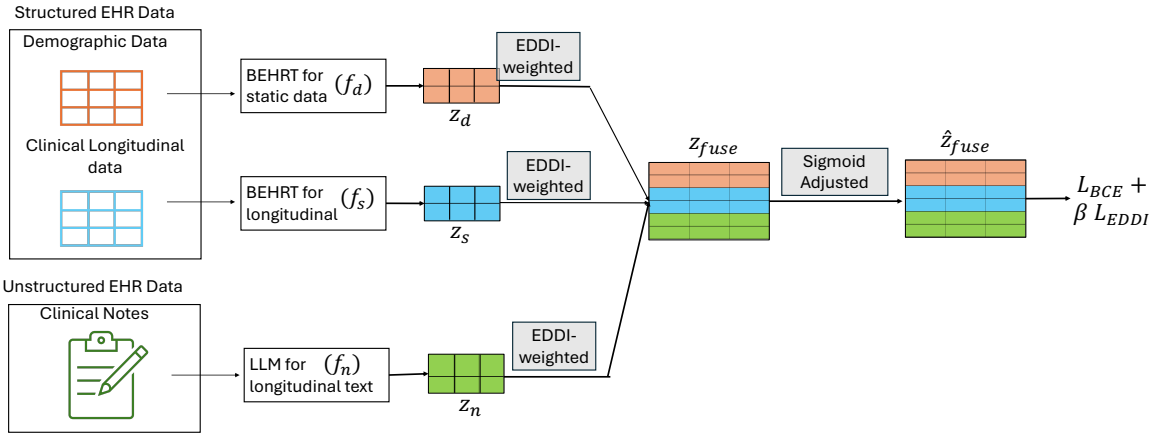

Figure 1: Model

We define this process more mathematically below, where each latent embedding is denoted as z and operations within the network are denoted by the function f. We define each model as:

$$z_m = f_m(X_m), \qquad z_{fuse} = f_{fuse}(z_1, \ldots, z_m), \qquad y = f_{cls}(z_{fuse})$$

where f_m is the m^{th} unimodal model, f_{fuse} is the fusion (aggregation) method used to fuse latent embeddings from all modalities, and f_{cls} maps the fused information to a binary class variable, y. Combining all the operations above, we can define our multimodal model f_{multi} as:

$$y = f_{multi}(X_1, \ldots, X_m) = f_{cls}\left(f_{fuse}\left(f_1(X_1), \ldots, f_m(X_m)\right)\right)$$

One simple example of a fusion method, f_{fuse}, could be taking the average of each modality:

$$z_{fuse} \rightarrow \frac{1}{M} \sum_{m=1}^{M} z_m \tag{1}$$

3.2. Fairness-aware Multimodal Fusion

In the joint fusion discussed in Equation 1, each modality is weighted equally towards fused embeddings z_{fuse}. We propose to weigh embeddings from each modality before combining them for multimodal fusion. Our goal is to find the relative weights of each modality towards the multimodal fusion based on its fairness performance.

$$z_{fuse}^{(t)} = \sum_{m=1}^{M} \hat{w}_m^{(t)} \cdot (z_m \cdot W_m) \tag{2}$$

where $z_{fuse}^{(t)}$ is the multimodality embedding at iteration t of training, and W_m is a trainable weight matrix that projects the m^{th} modality. Finally, $\hat{w}_m^{(t)}$ is a gating vector in t^{th} iteration

for the m^{th} modality that is influenced by fairness. This vector is defined more specifically below. Intuitively, this vector serves to scale dimensions within modalities that promote fairness and is recomputed at each iteration, t.

EDDI-weighted Fusion: To measure fairness for each modality, we compute the EDDI value across training iterations. For every modality we attach a light-weight classification head to its embedding z_m. These heads are excluded from the loss, so they receive no gradient updates and does not impact overall model training. The fixed probe heads used to estimate modality-level EDDI are described in Appendix A. EDDI across sub-groups can be combined by taking the mean across all subgroups, as proposed by Wang et al. (2024):

$$EDDI = \frac{1}{|S|} \sum_{s \in S} \frac{ER_s - OER}{\max(OER, 1 - OER)} \tag{3}$$

where ER_s defines the error rate for subgroup s (eg. White, Black, Asian) belonging to a sensitive attribute (eg. race) and OER is the overall error rate for the dataset (i.e., the expected error rate). Error rate is defined in the customary manner as the percentage of predicted class labels that do not match the ground truth label. A potential downside of equation 3 is that it adds all positive and negative EDDI values together, which can misrepresent the overall EDDI value for a sensitive attribute. To mitigate this, we propose to combine EDDI-values for all subgroups (eg. White, Black, Asian. etc) in a sensitive attribute (eg. race) using a sign-agnostic method. In this method, we take mean of the root of the sum of squares values of individual $EDDI_s$ for each subgroup. This helps ensure equitable fairness estimation across all demographic subgroups irrespective of their EDDI being positive or negative.

$$EDDI_s = \frac{ER_s - OER}{\max(OER, 1 - OER)} \tag{4}$$

$$EDDI_S = \frac{1}{|S|} \sqrt{\sum_{s \in S} EDDI_s^2} \tag{5}$$

We then take the mean of the $EDDI_S$ value across all sensitive attributes (like, race, age, insurance) to obtain overall EDDI across each modality. Based on this EDDI, we calculate the weight parameter w_m for each modality that modulates each embedding vector dimension through element-wise scaling (i.e., an element-wise scaling function). Embeddings from multiple modalities are adjusted using the following:

$$w_m^{(t)} = w_m^{(t-1)} + \gamma \cdot \left(\max_{m \in M} (EDDI_m^{(t)}) - EDDI_m^{(t)} \right) \tag{6}$$

where $w_m^{(t)}$ is the weight in t^{th} iteration, γ is the learning rate, and $EDDI_m^t$ is the mean EDDI value across all sensitive attributes for m modality. A lower value of EDDI reflects a group with more fair treatment of groups. γ controls how much $w_m^{(t)}$ should change each iteration based on the EDDI value for that iteration. The initial weight for each modality is assigned equally, where $w_m^{(0)} = \frac{1}{M}$. Therefore, when γ is 0 the result is that equal weight is given to each modality for all iterations, which reduces to average joint fusion given in Equation 1. By subtracting $EDDI_m$ from $\max(EDDI_m)$ we ensure that modality with

lower $EDDI_m$ receives more weight. Lastly, we take the normalized weight value across all modalities such that the sum of weights for all modalities is always unity.

$$\hat{w}_m^{(t)} = \frac{w_m^{(t)}}{\sum_{m \in M} w_m^{(t)}} \tag{7}$$

Sigmoid-weighted Feature Selection: As an extension to the proposed EDDI-weighted fusion, we also propose a feature gating mechanism to help promote more fair individual features within each modality. After obtaining the EDDI-weighted embedding for each modality $z_{mw} = \hat{w}_m \cdot (z_m \cdot W_m)$, we use a sigmoid activation to adjust embeddings by learning a weight parameter $\sigma(\cdot)$ that modulates each dimension of the embedding vector through element-wise scaling. For a given iteration, t, we first concatenate each $z_{mw}^{(t)}$ for all $m \in M$ into the vector z_{concat} and take the dot product of sigmoid layer weights with each unimodal embedding:

$$z_{concat}^{(t)} = \text{concat}(z_{1w}^{(t)}, \ldots, z_{mw}^{(t)}) \tag{8}$$

$$\hat{z^t} = \sigma(W) \odot z_{\text{concat}}^t \tag{9}$$

where $\hat{z}^t$ is the sigmoid adjusted embedding to modulate features from each modality and W is the trained weight vector,and z_{concat} is the 768-D vector obtained after concatenating 256-D vector from each modality. To incorporate into our final model, we simply replace z_{fuse} from Eq. 2 with $\hat{z^t}$ in the classification layer. We later use the sigmoid output to analyze which features from each modality are weighted more or less for the final prediction by aggregating weights for each 256 features in 768-D W vector.

3.3. Loss Functions

Joint fusion allows the gradient update to flow through all unimodal and fused models together to train the whole model on the desired outcome. We use a combination of binary-cross entropy and EDDI loss to optimize our fusion model f_{total}

$$\mathcal{L}_{total} = \mathcal{L}_{BCE} + \lambda \cdot \mathcal{L}_{EDDI} \tag{10}$$

where $\mathcal{L}_{BCE}$ is binary cross-entropy loss, $\mathcal{L}_{EDDI}$ is the mean of $EDDI_S$ (Equation 5) across all sensitive attributes, and λ is the hyperparameter to control the contribution of each loss.

3.4. Model for Multimodal Fusion

We will use different transformer-based models tailored for structured and unstructured data in our multimodal architecture.

BEHRT for Demographic and Longitudinal Structured Data: BEHRT (Li et al., 2020) is a transformer-based model specifically designed to process structured electronic health record (EHR) data, capturing the temporal, demographic, and clinical context of patient histories. It utilizes embedding layers for categorical values such as disease codes, age categories, and demographic data. These embeddings are combined with positional and segment embeddings to preserve temporal information about visits. We will utilize

Table 1: Outcome Label Distribution

Outcome	Total Patients	Positive Cases	Percentage (%)
Short-Term Mortality	33,721	3,417	10.13
Length of Stay(LOS)	33,721	4,991	14.80
Mechanical Ventilation	33,721	30,356	90.02

BEHRT model in our code for structured clinical and demographic data. We will extend the BEHRT model to include numerical values from patients' labs and vitals. For our joint fusion method, we will use $[CLS]$ embedding from last layer as the representative embedding for each patient's structured clinical and demographic data.

BioClinicalBERT for Unstructured Text Data: We will use BioClinicalBERT (Alsentzer et al., 2019) to process unstructured clinical data from electronic health records (EHR). We began by preprocessing the dataset, where the text data underwent tokenization using the BioClinicalBERT tokenizer, followed by embedding extraction via a fine-tuned BioClinicalBERT model. Again we will use [CLS] embedding from last layer as patient-level embeddings for use in joint-fusion framework.

Multimodal Fusion: We will combine embeddings from three data modalities — demographic (z_d), longitudinal structured clinical (z_s), and unstructured clinical notes (z_n) — to implement our fairness-aware joint fusion framework. We will bring embeddings from all three unimodals into same $n - dimensional$ before applying fairness-aware weighting.

4. Cohort

4.1. Cohort Selection

Our cohort selection steps followed the benchmark by Wang et al. (2020) where we used only first ICU visit for all patients above 15 years of age. In all cases, we use the first 24 hours of a patient's data, only considering patients with at least 30 hours of present data. This 6 hour gap time is critical to prevent temporal label leakage. For structured data we followed Feature Set C from Purushotham et al. (2018) to select 136 features from MIMIC-III which is a superset of features selected by Wang et al. (2020). We aggregated temporal features into 2-hour bins. Additionally, textual embeddings from BioClinicalBERT were incorporated to capture information from physician notes, nursing progress records, and radiology reports. (see Appendix B for full preprocessing details).

4.2. Prediction Tasks

We use our data to predict three binary classification tasks: In-ICU Mortality, Length of Stay (LOS) ≥ 7, and Mechanical Ventilation.

The In-ICU Mortality label was extracted from the DEATHTIME field in the MIMIC-III dataset, where a non-null entry signifies a death event. The length of stay (LOS) prediction task aims to classify whether a patient's ICU stay will exceed three days using the first 24 hours of clinical data. The label was derived from the INTIME and OUTTIME timestamps in the ICUSTAYS table of the MIMIC-III dataset. Mechanical ventilation is a critical intervention in ICU patients, often indicating severe respiratory distress or failure. We identified

Table 2: Sensitive Attribute Distribution

Feature	Subcategory	Count	Percentage (%)
Age Bucket	15–29	1,832	5.4
	30–49	5,729	17.0
	50–69	13,344	39.6
	70+	12,816	38.0
Ethnicity	White	23,887	70.8
	Black	2,567	7.6
	Hispanic	1,076	3.2
	Asian	670	2.0
	Other	5,521	16.4
Insurance	Medicare	17,163	50.9
	Private	12,151	36.0
	Medicaid	2,889	8.6
	Government	1,060	3.1
	Self Pay	458	1.4

ventilation-related item IDs from the CHARTEVENTS and PROCEDUREEVENTS_MV tables, including ventilator settings, oxygen therapy usage, and extubation events. Additionally, procedural records indicating intubation or tracheostomy were incorporated to refine label accuracy.(full derivation in Appendix C).

4.3. Sensitive Attributes

We will use sensitive attributes - ethnicity (in the MIMIC dataset, both race and ethnicity are recorded under ethnicity), insurance, and age. Each attribute is further divided into subgroups with ethnicity having 5 subgroups - White, Black, Hispanic, Asian and Other, insurance which can act as a proxy for socio-economic status has 5 subgroups - Medicare, Private, Medicaid, Goverment, and Self-pay, and age divided into subgroups - 15-29, 30-49, 50-69, and 70+. We have shared Table 2, which shows the distribution of all subgroups in the sensitive attributes.

5. Experiments and Results

5.1. Implementation Details

To implement multimodal models we project the [CLS] embeddings from all three modalities into a shared 256-dimensional space. In the initial iteration, each modality is assigned equal weight, and their embeddings are concatenated to form a 768-dimensional representation. In subsequent iterations, we use the mean EDDI weights computed across three sensitive attributes—ethnicity, insurance status, and age—to weight each modality. These EDDI weights are updated at each iteration using a learning rate (γ) of 0.5. The final concatenated embedding is passed through a linear classification layer with a hidden dimension of 512 and a dropout rate of 0.2 to predict three binary outcomes.

To address class imbalance in the prediction tasks, we use a weighted binary cross-entropy loss function, where class weights are determined via the Inverse of Number of Samples (INS) method. Specifically, we employ PyTorch's BCEWithLogitsLoss, assigning higher weights to minority classes to mitigate skewed distributions.

The model is optimized using the AdamW optimizer with a learning rate of 1×10^{-5} and a weight decay of 0.01 for regularization. We also incorporate the ReduceLROnPlateau scheduler from to adaptively reduce the learning rate based on validation performance.

Training is performed on an 80-20 split of the data, with 80% used for training and 20% for testing. Additionally, 5% of the training data is held out for validation. Early stopping is applied with a patience of 5 epochs, terminating training if validation loss does not improve across 5 consecutive epochs. The model checkpoint with the lowest validation loss is selected for final evaluation on the test set. All hyper-parameters were tuned via grid search on the validation split; see Appendix D for full details. Code is available on GitHub.

5.2. Baseline Comparison

5.3. Baseline Comparison

We compare our proposed model against several state-of-the-art methods designed to mitigate bias in EHR prediction tasks. Evaluations are conducted on both predictive performance and fairness metrics to provide a comprehensive assessment of model effectiveness.

Demographic-free Classification (DfC): It is an established method to reduce the impact of sensitive attributes on outcome is by unawareness. DfC is based on the concept of unawareness, which states that if demographic features are not included in input data, it should have minimal effect on output.

AdvDebias (Zhang et al., 2018; Yang et al., 2023): It uses adversarial training to debias patient representations. It trains an adversary model such that the model cannot correctly classify patient's sensitive attributes from the patient representations, thus debiasing patient representations. A classifier is trained on debiased patient representation to improve fairness.

Fair Patient Model (FPM) (Sivarajkumar et al., 2023): FPM employs a Stacked Denoising Autoencoder and a weighted reconstruction loss for equitable patient representations.

FairEHR-CLP (Wang et al., 2024): FairEHR-CLP uses contrastive learning between patient data and synthetically generated counterfactual samples with different sensitive attributes but the same medical histories as the original sample. By reducing contrastive loss between original and counterfactuals, the aim is to reduce bias in the multimodal EHR model.

5.4. Ablation Analysis

As part of the ablation analysis, we will compare our model with below modifications:

BEHRT: We will use unimodal BEHRT only for structured clinical data modality (X_s) to predict three prediction tasks.

BioClinicalBERT: We will use unimodal BioClinicalBERT only for unstructured clinical text modality (X_n) to predict three prediction tasks.

Average Fusion: Multimodal model with three modalities combined using Eq. 1

Table 3: Baseline comparison using performance and fairness evaluation across three prediction tasks. We report average results over 5 runs with std deviation. EDDI, and EO are averaged over three sensitive attributes. Bold indicates best results.

Task	Model	AUROC (↑)	AUPRC (↑)	EDDI(%) (↓)	EO(%)(↓)
In-ICU Mortality	DfC	$0.90_{\pm 0.02}$	$0.71_{\pm 0.02}$	$0.79_{\pm 0.08}$	$5.80_{\pm 1.02}$
	AdvBias	$0.93_{\pm 0.01}$	$0.75_{\pm 0.01}$	$2.40_{\pm 1.05}$	$9.91_{\pm 2.01}$
	FPM	$0.93_{\pm 0.02}$	$0.74_{\pm 0.01}$	$6.94_{\pm 1.01}$	$14.10_{\pm 2.64}$
	FairEHR-CLP	$0.92_{\pm 0.01}$	$0.74_{\pm 0.01}$	$8.84_{\pm 0.66}$	$16.10_{\pm 2.32}$
	FAME (Ours)	$\mathbf{0.94_{\pm 0.02}}$	$\mathbf{0.82_{\pm 0.01}}$	$\mathbf{0.44_{\pm 0.04}}$	$\mathbf{4.25_{\pm 1.72}}$
LOS ≥ 7	DfC	$0.98_{\pm 0.02}$	$0.91_{\pm 0.01}$	$0.55_{\pm 0.01}$	$2.83_{\pm 1.20}$
	AdvBias	$0.96_{\pm 0.01}$	$0.85_{\pm 0.01}$	$2.39_{\pm 1.09}$	$5.70_{\pm 1.56}$
	FPM	$0.96_{\pm 0.01}$	$0.86_{\pm 0.05}$	$3.43_{\pm 1.58}$	$11.10_{\pm 2.05}$
	FairEHR-CLP	$0.96_{\pm 0.01}$	$0.85_{\pm 0.03}$	$3.85_{\pm 0.09}$	$12.75_{\pm 1.28}$
	FAME (ours)	$\mathbf{1.00_{\pm 0.02}}$	$\mathbf{1.00_{\pm 0.02}}$	$\mathbf{0.02_{\pm 0.00}}$	$\mathbf{0.06_{\pm 0.01}}$
Mechanical Ventilation	DfC	$0.78_{\pm 0.03}$	$0.97_{\pm 0.03}$	$\mathbf{1.29_{\pm 0.08}}$	$2.58_{\pm 1.21}$
	AdvBias	$0.84_{\pm 0.02}$	$0.97_{\pm 0.03}$	$1.73_{\pm 0.08}$	$15.70_{\pm 3.15}$
	FPM	$0.83_{\pm 0.02}$	$0.97_{\pm 0.02}$	$7.51_{\pm 1.65}$	$13.50_{\pm 2.97}$
	FairEHR-CLP	$0.84_{\pm 0.01}$	$0.97_{\pm 0.02}$	$9.67_{\pm 1.37}$	$16.40_{\pm 2.11}$
	FAME (Ours)	$\mathbf{0.84_{\pm 0.02}}$	$\mathbf{0.97_{\pm 0.02}}$	$2.77_{\pm 1.14}$	$\mathbf{0.55_{\pm 2.01}}$

Sigmoid-only: Our proposed multimodal model with three modalities combined using concatenated fusion method in Eq. 8 and sigmoid function in Eq. 9.

EDDI-only: Our proposed multimodal model with three modalities combined using Eq. 2 with EDDI-weights from Eq. 6 and Eq. 7 and no Sigmoid-weighted feature selection.

5.5. Evaluation Metrics

We evaluate the model using both standard predictive performance metrics and fairness-aware assessments. We use Area Under the Receiver Operating Characteristic Curve (AUROC) and Area Under the Precision-Recall Curve (AUPRC) to assess overall model performance and precision-recall trade-offs. To assess fairness across different demographic groups, we employed EDDI and Equalized Opportunity (EO). To calculate EO we took mean of aggregated True positive rate (TPR) and False positive rate (FPR). To calculate aggregated TPR and FPR, we first calculate the absolute pairwise difference of TPR and FPR between each subgroup of a sensitive attribute and then take the mean over three attributes (Wang et al., 2024). We also report EDDI by first aggregating at subgroup-level (Eq. 5) and then taking the arithmetic mean over all sensitive attributes. All these fairness metrics need to be lower to demonstrate the fairness of a model.

Table 4: Ablation analysis across modalities and model components using performance and fairness evaluation across three prediction tasks. We report average results over 5 runs with std deviation. EDDI, and EO are averaged over three sensitive attributes. Bold indicates best results.

Task	Model	AUROC (↑)	AUPRC (↑)	EDDI(%) (↓)	EO(%)(↓)
In-ICU Mortality	BEHRT	$0.83_{\pm 0.02}$	$0.40_{\pm 0.03}$	$2.08_{\pm 0.08}$	$5.20_{\pm 1.97}$
	BioClinicalBERT	$0.93_{\pm 0.03}$	$0.78_{\pm 0.03}$	$0.82_{\pm 0.04}$	$4.80_{\pm 2.13}$
	Average Fusion	$0.93_{\pm 0.03}$	$0.74_{\pm 0.03}$	$1.39_{\pm 1.45}$	$6.91_{\pm 2.86}$
	Sigmoid-only	$0.86_{\pm 0.02}$	$0.40_{\pm 0.02}$	$6.82_{\pm 1.32}$	$15.53_{\pm 3.63}$
	EDDI-only	$0.92_{\pm 0.03}$	$0.69_{\pm 0.03}$	$1.66_{\pm 0.06}$	$7.00_{\pm 3.29}$
	FAME (Ours)	$\mathbf{0.94_{\pm 0.02}}$	$\mathbf{0.82_{\pm 0.01}}$	$\mathbf{0.44_{\pm 0.04}}$	$\mathbf{4.25_{\pm 1.72}}$
LOS ≥ 7	BEHRT	$0.76_{\pm 0.02}$	$0.35_{\pm 0.01}$	$1.34_{\pm 0.07}$	$4.24_{\pm 1.59}$
	BioClinicalBERT	$0.96_{\pm 0.03}$	$0.85_{\pm 0.02}$	$0.66_{\pm 0.05}$	$4.01_{\pm 0.07}$
	Average Fusion	$0.96_{\pm 0.03}$	$0.85_{\pm 0.03}$	$0.34_{\pm 0.02}$	$3.97_{\pm 1.05}$
	Sigmoid-only	$1.00_{\pm 0.02}$	$0.97_{\pm 0.02}$	$0.09_{\pm 0.02}$	$0.20_{\pm 0.01}$
	EDDI-only	$1.00_{\pm 0.02}$	$0.99_{\pm 0.02}$	$0.07_{\pm 0.02}$	$0.06_{\pm 0.01}$
	FAME (ours)	$\mathbf{1.00_{\pm 0.02}}$	$\mathbf{1.00_{\pm 0.02}}$	$\mathbf{0.02_{\pm 0.00}}$	$\mathbf{0.06_{\pm 0.01}}$
Mechanical Ventilation	BEHRT	$0.80_{\pm 0.03}$	$0.97_{\pm 0.02}$	$4.72_{\pm 1.77}$	$8.23_{\pm 3.43}$
	BioClinicalBERT	$0.78_{\pm 0.03}$	$0.97_{\pm 0.03}$	$1.98_{\pm 1.16}$	$2.75_{\pm 1.70}$
	Average Fusion	$0.83_{\pm 0.02}$	$0.97_{\pm 0.03}$	$3.11_{\pm 2.21}$	$6.93_{\pm 3.15}$
	Sigmoid-only	$0.67_{\pm 0.03}$	$0.94_{\pm 0.02}$	$4.59_{\pm 2.58}$	$8.79_{\pm 3.25}$
	EDDI-only	$\mathbf{0.88_{\pm 0.02}}$	$\mathbf{0.98_{\pm 0.02}}$	$4.96_{\pm 1.88}$	$7.17_{\pm 2.78}$
	FAME (Ours)	$0.84_{\pm 0.02}$	$0.97_{\pm 0.02}$	$\mathbf{2.77_{\pm 1.14}}$	$\mathbf{0.55_{\pm 2.01}}$

6. Results

6.1. Baseline Comparison

Table 3 presents a comparative evaluation of our proposed model against several state-of-the-art baselines using four key metrics: AUROC, AUPRC, Error Distribution Disparity Index (EDDI), and Equalized Odds (EO). Although DfC consistently underperforms in AUROC and AUPRC, it achieves lower bias compared to other baselines. This indicates that completely omitting demographic features can help reduce bias, but often at the significant cost of predictive performance. These findings suggest that rather than exclusion, a more effective strategy may be to regulate the influence of demographic features to strike a balance between fairness and accuracy. Our proposed model excels in AUROC and AUPRC compared to all baselines and has reduced bias compared to baselines in most settings. It shows the importance of including all modalities to improve performance but also controlling their contribution to produce equitable outcomes.

6.2. Ablation Analysis:

Modality Analysis: As part of our ablation study in Table 4, we first evaluate the predictive effectiveness of individual modalities by training two unimodal models: BEHRT, which uses only structured clinical data, and BioClinicalBERT, which relies solely on clini-

Table 5: Fairness Results for all sensitive attributes compared across different components of our proposed model.

Model Model	EDDI (%) (↓) (Age)	EDDI(%)(↓) (Ethnicity)	EDDI (%) (↓) (Insurance)	EO (%)(↓) (Age)	EO (%)(↓) (Ethnicity)	EO (%)(↓) (Insurance)
Sigmoid-Only	1.93	9.06	1.89	3.35	19.91	4.18
EDDI-only	**0.99**	5.31	**1.04**	**1.34**	10.89	**1.99**
FAME	1.85	**0.47**	1.10	1.57	**0.8**	2.48

cal notes. BioClinicalBERT outperforms BEHRT by 12% in AUROC and 77% in AUPRC and also reduces bias by lower EDDI by 56% and EO by 26%, highlighting the richness and diversity of information embedded in clinical text. Clinical notes often capture not only clinical observations but also implicit demographic and social context, contributing to their predictive strength. When evaluated for fairness, our proposed model shows lower bias in most settings, over BioClinicalBERT (best performing unimodal), demonstrating the benefit of leveraging complementary information across modalities.

Component Analysis: Our ablation analysis of model components (Table 4) reveals that our proposed model FAME, which incorporates EDDI-based modality weighting, outperform the average fusion baseline that weighs all modalities equally. Specifically, FAME achieves overall improvements of 3% in AUROC, and 9% in AUPRC and significant bias reductions of 57% and 76% in EDDI and EO, highlighting the effectiveness of EDDI-weighted modality aggregation. Furthermore, the Sigmoid-only model exhibits a greater increase in bias compared to the EDDI-only model when both are evaluated against the full model, reinforcing the importance of EDDI-based weighting. The full model consistently yields the best fairness outcomes (lowest EDDI scores), suggesting that the combination of EDDI-weighted aggregation and Sigmoid-based feature selection provides complementary advantages for both performance and fairness.

6.3. Fairness Comparison Across Sensitive Attributes:

In Table 5 we compare EDDI, and EO separately for our three sensitive attributes across different variations of our model. We average EDDI, and EO across three tasks. We observe that EDDI-only and FAME has consistently has lower EDDI and EO across all sensitive attributes compared to Sigmoid-only, reinforcing the importance of EDDI-weighting of individual modalities. Out of three sensitive attributes age and insurance (a proxy for socio-economic status) are more biased compared to ethnicity (also includes race).

6.4. Sensitivity Analysis

We conducted a sensitivity analysis of the hyperparameter λ (Eq. 10) to examine the trade-off between the binary cross-entropy loss and the EDDI loss in shaping model performance. Figure 2 illustrates a general trend in which both EDDI and EO decreases as AUPRC increases, underscoring the inherent tension between fairness and accuracy. Performance follows a U-shaped curve with respect to λ: where an intermediate value (here, $\lambda = 0.8$) yields the best overall trade-off.

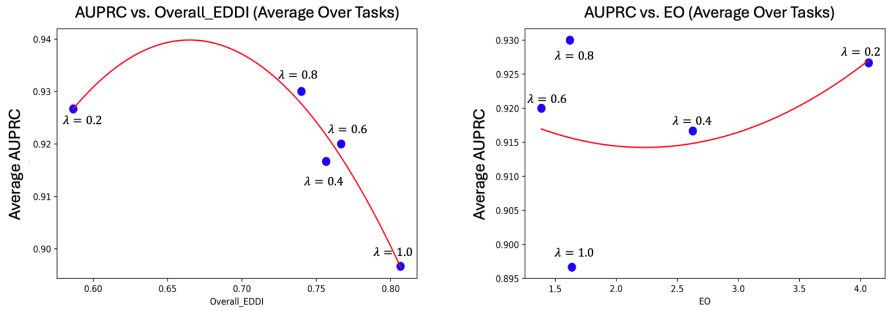

Figure 2: Sensitivity analysis of the effect of λ on performance and fairness. We compare AUPRC vs. EDDI and AUPRC vs. EO, averaged over all tasks.

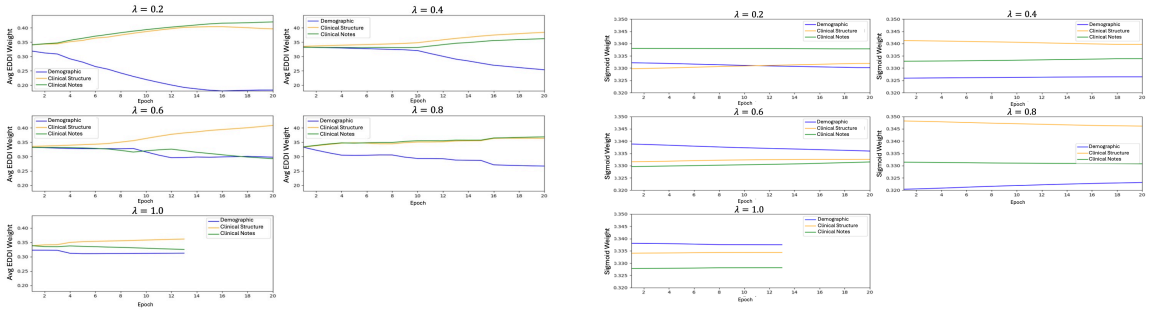

Figure 3: Visualization of EDDI and Sigmoid weights over training iterations, averaged across three tasks. Shown for different λ values.

This trend in Figure 2 is not strictly monotonic, a behaviour often seen in deep networks because of regularisation effects. To provide further clarity on the trend, which shows that $\lambda = 0.8$ yields the best results, we draw on evidence presented in the next section. We discuss in detail the interaction between λ, EDDI, and Sigmoid weights in Section 6.6, offering a deeper understanding of how λ influences the EDDI and Sigmoid weights learned during the training affecting the overall performance shown in Figure 2.

6.5. EDDI-weights and Sigmoid-weights Visualization:

Our proposed model is trained on joint loss of BCE and EDDI. To investigate how EDDI weight and Sigmoid feature selection evolve while our model is being optimized on joint loss we visualize changing weights over training epochs. In the left panel of Figure 3 we illustrate how EDDI weights, computed using Eq.6, evolve during training. The right panel shows the progression of Sigmoid weights, derived from Eq.9, over the course of model optimization. Both EDDI and Sigmoid weights are averaged across three tasks for

visualization. Additionally, the Sigmoid layer weights are averaged over the 256-dimensional vectors corresponding to each modality, which are concatenated as described in Eq. 8.

The EDDI weights, initialized uniformly across modalities, gradually diverge during training, assigning higher weights to the structured clinical and unstructured text modalities, while the weight for the demographic modality consistently decreases across all settings. This trend is consistent with our baseline comparison, where removing the demographic modality from the DfC model improves EDDI performance. The Sigmoid feature selector continues to extract features from the demographic modality but assigns them lower importance compared to the other two modalities. EDDI weights for both structured and unstructured data increase over time, with the text modality receiving slightly higher weights in some settings (λ=0.2, 0.8). The Sigmoid layer, however, assigns greater emphasis to features from the structured modality (λ=0.4, 0.8). At the optimal λ value of 0.8 from our sensitivity analysis in Figure 2, EDDI assigns nearly equal weights to the structured and text modalities, while the feature selection layer slightly favors the structured modality. Also, the anomaly behaviors at λ=1.0 for EDDI in Figure 2 can be attributed to higher demographic weight in the right panel of Figure 3.

6.6. Interactions between λ, EDDI weights and Sigmoid weights

In the left panel of Figure 2, we observe that $\lambda = 1.0$ yields the lowest AUPRC, as expected, and $\lambda = 0.2$ achieves higher AUPRC than $\lambda = 0.4$ and $\lambda = 0.6$, which aligns with our expectations. Interestingly, $\lambda = 0.8$ gives the highest AUPRC, though only marginally better than $\lambda = 0.2$. On the fairness side, while $\lambda = 1.0$ results in the worst fairness score, $\lambda = 0.2$ yields the best fairness outcome, which deviates from the expected trend. Nevertheless, $\lambda = 0.8$ performs better in terms of EDDI compared to $\lambda = 0.4$ and $\lambda = 0.6$, supporting the overall pattern. Looking at Figure 3 for further insight, we see that $\lambda = 1.0$ assigns nearly equal EDDI weights across all three modalities, similar to the initial EDDI weights, suggesting that the model is underfitting and not effectively learning meaningful representations. The right panel of Figure 3 also shows disproportionately high feature selection from the demographic modality, with weights again remaining close to the initial values assigned in the 0th iteration at $\lambda = 1.0$, further indicating poor learning behavior. This underfitting leads to suboptimal classification performance and contributes to the anomalous fairness outcomes observed.

In the right panel of Figure 2, AUPRC for $\lambda = 1.0$ is the lowest, as expected, with $\lambda = 0.2$ achieving higher AUPRC than $\lambda = 0.6$ and $\lambda = 0.4$, which aligns with expected behavior. The difference in AUPRC between $\lambda = 0.4$ and $\lambda = 0.6$ is minimal. On fairness metric $\lambda = 1.0$, $\lambda = 0.8$, and $\lambda = 0.6$ yield lower EO compared to $\lambda = 0.4$, and $\lambda = 0.2$. Notably, $\lambda = 0.8$ delivers the best overall performance in terms of both EO and AUPRC.

Further supporting this, Figure 3 shows that $\lambda = 0.8$ starts learning weights for each modality early in the training and results in higher EDDI weights for the structured and unstructured modalities compared to demographic, and it learns different weights for each modality giving higher weights to structured modality and least weight to demographic. Both observations are consistent with the strong performance observed at $\lambda = 0.8$.

7. Discussion

The improved fairness and prediction performance results achieved by FAME highlight the limitations of relying on a single modality, even one as rich as clinical notes, and emphasize the value of integrating multiple modalities to capture complementary signals and mitigate bias. These findings reinforce the importance of explicitly accounting for modality-specific fairness contributions, as done through EDDI-based weighting. Compared to naive equal-weighted fusion, EDDI-weighted aggregation leads to more equitable outcomes across subgroups, as evidenced by reduced EDDI and EO scores.

Interestingly, while the model continues to access demographic information, minimizing its influence appears to support better fairness. The consistent reduction in demographic weights, observed through both EDDI and Sigmoid components, suggests that while such features may offer signal, their overemphasis can amplify bias. In contrast, structured and unstructured clinical data show increasing importance over training, with their divergent selection dynamics indicating complementary strengths of our model components.

The relatively higher weighting of the structured modality suggests a promising direction for future work: extracting both clinical and non-clinical information from unstructured text and integrating it into structured (tabular) formats. Such an approach could enhance not only model performance but also fairness by making valuable information from clinical notes more accessible to structured modeling techniques.

Limitations and Future Work While our study demonstrates the potential of fairness-aware multimodal fusion, there are a few limitations worth noting. First, we did not incorporate image modalities in our current experiments. Although the proposed framework is designed to support additional modalities, our evaluation was focused on structured, unstructured text, and demographic data. Second, our analysis considered a limited set of sensitive attributes—race, insurance status, and age—which may not fully represent the range of factors that contribute to disparities in healthcare. Lastly, the results are based on a single dataset, which may affect the generalizability of our findings across datasets.

As part of future work, we plan to extend our evaluation to datasets that include imaging data to better understand how visual information complements other modalities in enhancing both performance and fairness. We also aim to explore a broader range of sensitive attributes, including Social Determinants of Health (SDoH) identified through clinical notes, to capture a more complete picture of patient subgroups. Finally, we hope to validate our approach across multiple datasets to further examine how dataset characteristics influence model behavior and fairness outcomes.

Acknowledgments

This research was supported by Faculty Research Acceleration Grant, graciously provided by the SMU O'Donnell Data Science and Research Computing Institute.

References

Ankita Agarwal, Tanvi Banerjee, William L Romine, and Mia Cajita. Debias-clr: A contrastive learning based debiasing method for algorithmic fairness in healthcare applica-

tions. In *2024 IEEE International Conference on Big Data (BigData)*, pages 6411–6419. IEEE, 2024.

Ritu Agarwal, Margret Bjarnadottir, Lauren Rhue, Michelle Dugas, Kenyon Crowley, Jessica Clark, and Gordon Gao. Addressing algorithmic bias and the perpetuation of health inequities: An ai bias aware framework. *Health Policy and Technology*, 12(1):100702, 2023.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

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.

Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1):4209, 2022.

Hejie Cui, Xinyu Fang, Ran Xu, Xuan Kan, Joyce C Ho, and Carl Yang. Multimodal fusion of ehr in structures and semantics: Integrating clinical records and notes with hypergraph and llm. *arXiv preprint arXiv:2403.08818*, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Kexin Huang, Abhishek Singh, Sitong Chen, Edward T Moseley, Chih-Ying Deng, Naomi George, and Charlotta Lindvall. Clinical xlnet: Modeling sequential clinical notes and predicting prolonged mechanical ventilation. *arXiv preprint arXiv:1912.11975*, 2019.

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, 2020a.

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, 2020b.

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.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):7155, 2020.

Zheng Liu, Xiaohan Li, and Philip Yu. Mitigating health disparities in ehr via deconfounder. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–6, 2022.

Weimin Lyu, Xinyu Dong, Rachel Wong, Songzhu Zheng, Kayley Abell-Hart, Fusheng Wang, and Chao Chen. A multimodal transformer: Fusing clinical notes with structured ehr data for interpretable in-hospital mortality prediction. In *AMIA Annual Symposium Proceedings*, volume 2022, page 719, 2023.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54 (6):1–35, 2021.

Stephen R Pfohl, Agata Foryciarz, and Nigam H Shah. An empirical characterization of fair machine learning for clinical risk prediction. *Journal of biomedical informatics*, 113: 103621, 2021.

Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmarking deep learning models on large healthcare datasets. *Journal of biomedical informatics*, 83:112–134, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67, 2020.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2020, page 2359, 2020.

Reva Schwartz, Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. *Towards a standard for identifying and managing bias in artificial intelligence*, volume 3. US Department of Commerce, National Institute of Standards and Technology, 2022.

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.

Raj Sharman. Data challenges and societal impacts–the case in favor of the blueprint for an ai bill of rights (keynote remarks). In *International Conference on Big Data Analytics*, pages 3–15. Springer, 2022.

Sonish Sivarajkumar, Yufei Huang, and Yanshan Wang. Fair patient model: Mitigating bias in the patient representation learned from the electronic health records. *Journal of biomedical informatics*, 148:104544, 2023.

Yao-Hung Hubert Tsai, Martin Q Ma, Muqiao Yang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Multimodal routing: Improving local and global interpretability of multimodal language analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2020, page 1823. NIH Public Access, 2020.

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.

Shirly Wang, Matthew BA McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C Hughes, and Tristan Naumann. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM conference on health, inference, and learning*, pages 222–235, 2020.

Yuqing Wang, Malvika Pillai, Yun Zhao, Catherine Curtin, and Tina Hernandez-Boussard. Fairehr-clp: Towards fairness-aware clinical predictions with contrastive learning in multimodal electronic health records. *arXiv preprint arXiv:2402.00955*, 2024.

Jenny Yang, Andrew AS Soltan, David W Eyre, Yang Yang, and David A Clifton. An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *NPJ digital medicine*, 6(1):55, 2023.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

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 (2021). *URL https://arxiv. org/abs/2111.11665. Lemma A*, 7, 2021.

Appendix A. Probe Heads for Modality-Level EDDI Computation

For every modality we attach a light-weight classification head—a single linear layer followed by a sigmoid—to its embedding z_m. These heads are *not optimised*: their outputs are excluded from the loss, so they receive no gradient updates and remain effectively fixed. At the end of each epoch we run them in `torch.no_grad()` mode on the training data to obtain predictions $\hat{y}_m$, from which subgroup error rates ER_s and the modality-level $\mathrm{EDDI}_m^{(t)}$ are computed. Thus they function solely as fairness probes supplying the $\mathrm{EDDI}_m^{(t)}$ values used in Eq. (6).

Appendix B. Data Preprocessing

Preprocessing involved structured and unstructured data extracted from the MIMIC-III database. Each data type required tailored techniques to prepare it for model training while addressing the specific challenges inherent in clinical datasets.

B.1. Structured Data Preprocessing

For the structured data, we load core MIMIC-III datasets (ADMISSIONS, PATIENTS, and ICUSTAYS) and convert critical timestamp fields (e.g., ADMITTIME, DISCHTIME, DEATHTIME, INTIME, OUTTIME) to datetime format. Columns are renamed for consistency and ICU stays are merged with admissions and patient demographics to provide comprehensive clinical context, including outcomes such as short-term mortality (based on the presence of DEATHTIME) and hospital readmission within 30 days (computed by assessing the interval between consecutive ICU admissions). Age at ICU admission is calculated from date of birth and categorized into predefined buckets (15–29, 30–49, 50–69, 70–89), while ethnicity and insurance information are standardized using rule-based mappings. Categorical features (except gender) are one-hot encoded to facilitate model input. Finally, the structured dataset is filtered to retain only patients whose IDs are common with the unstructured data.

B.2. Unstructured Data Preprocessing

For the unstructured data, we process the NOTEEVENTS dataset by converting CHARTDATE to datetime and applying text cleaning procedures to remove extraneous characters, standardize abbreviations, and eliminate unnecessary whitespace. Concatenating notes aggregate clinical notes corresponding to each patient's first ICU stay for a given hospital admission, and then split into chunks of approximately 512 tokens for compatibility with transformer-based models. The unstructured dataset is filtered to include only patients that overlap with the structured dataset. This integrated preprocessing pipeline ensures that structured (demographic, clinical, and laboratory) and unstructured (free-text clinical notes) data are aligned at the patient level, enabling a robust multimodal analysis.

Appendix C. Mechanical ventilation Labels

Mechanical ventilation was identified by integrating data from the CHARTEVENTS and PROCEDUREEVENTS_MV datasets. For CHARTEVENTS, we load the columns ICUS-

TAY_ID, CHARTTIME, ITEMID, VALUE, and ERROR, filter out rows with null values or errors, and retain only those with predefined ventilation-related ITEMIDs (e.g., 720, 223848, 223849, 467). A helper function (determine_flags) sets flags (mechvent, oxygentherapy, extubated, selfextubated) based on the recorded values. Similarly, the PROCEDUREEVENTS_MV dataset is filtered for extubation-related ITEMIDs (after renaming STARTTIME to CHARTTIME) to mark extubation events, with extubated set to 1 and other flags set to 0. These signals are concatenated and merged with ICU stay data from ICUSTAYS (adding SUBJECT_ID and HADM_ID), and then aggregated by taking the maximum value across the flags for each ICU stay. The final binary label is computed as:

$$mechanical\_ventilation = \max(mechvent, oxygentherapy, extubated, selfextubated)$$

A patient is labeled as 1 if any of these flags is positive during their ICU stay.

Appendix D. Hyper-parameter Search Details

All key hyper-parameters were chosen via grid search on the validation split, jointly optimising predictive performance (binary cross-entropy) and fairness (EDDI). The ranges explored and the final choices are summarised in Table 6.

- **Learning rate** $\in \{1 \times 10^{-5}, 5 \times 10^{-6}, 2 \times 10^{-5}\}$ and **batch size** $\in \{8, 16, 32\}$ were tuned first to ensure stable transformer and fusion-layer training. A learning rate of 1×10^{-5} with batch size 16 produced the best convergence and validation metrics.
- **Fairness weight** $\lambda \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ controls the EDDI loss term (Fig. 2, Sec. 6.4). $\lambda = 0.8$ gave the optimal AUPRC–vs.–fairness trade-off.
- **Weight-update rate** $\gamma \in \{0.5, 1.0\}$ (Eq. 6) sets the step size for modality weights (clipped at ± 0.05 per epoch). $\gamma = 1.0$ yielded the most consistent fairness gains without oscillation.
- **L1 regularisation** of the sigmoid gates used $\alpha \in \{0.01, 0.1\}$; $\alpha = 0.01$ encouraged modest sparsity without hurting accuracy.

Hyper-parameter	**Search grid (bold = chosen)**
Learning rate	5×10^{-6}, $\mathbf{1 \times 10^{-5}}$, 2×10^{-5}
Batch size	8, **16**, 32
λ (EDDI loss)	0.2, 0.4, 0.6, **0.8**, 1.0
γ (update rule)	0.5, **1.0**
L1 coefficient	**0.01**, 0.1

Table 6: Grid-search ranges and selected hyper-parameters.