TRIHAR: HIERARCHICAL ADAPTIVE ROUTING OF REPRESENTATIONS, MODALITIES, AND CROSS-MODAL INTERACTIONS FOR MULTIMODAL CLINICAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Conventional multimodal fusion methods typically rely on fixed encoder outputs, overlooking that informative representation depths may vary across modalities and patients. They also inadequately capture patient-specific variations in direct modality and cross-modal interaction contributions. We propose TriHAR, a hierarchical adaptive routing framework that integrates EHR, ECG, and CXR data through three routing processes. TriHAR selects an informative encoder depth within each modality and independently routes the resulting modality-specific predictions and EHR-centered cross-modal interaction predictions. Experiments on ICU length-of-stay prediction across 27 encoder combinations show that TriHAR achieves the best average rank for Accuracy, AUROC, and Precision and remains effective when either ECG or CXR is unavailable. Further analyses reveal that preferred representation depths vary across modalities and encoder architectures, while modality-expert and interaction-expert routing exhibit distinct yet complementary prediction characteristics. These results demonstrate the effectiveness of hierarchical adaptive routing across heterogeneous multimodal clinical representations.

1 INTRODUCTION

Multimodal clinical prediction combines heterogeneous patient data, including electronic health records (EHR), electrocardiograms (ECG), and chest X-rays (CXR) Adeyi et al. (2026); Rasekh et al. (2024). These modalities provide complementary clinical information: EHR data capture demographics, laboratory results, and physiological measurements Shickel et al. (2017); ECG signals reflect cardiac electrical activity; and CXR images reveal anatomical and pathological findings Berkaya et al. (2018); Çallı et al. (2021). Their integration can provide a more comprehensive assessment than any single modality. However, effective fusion remains challenging because the modalities differ substantially in structure, temporal resolution, dimensionality, and clinical meaning Hemker et al. (2024); Liu et al. (2025).

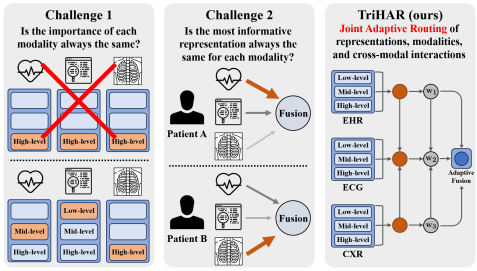


Figure 1: Two main challenges in multimodal clinical prediction addressed by TriHAR.

Most existing methods use the final output of each modality encoder and combine the resulting representations through concatenation, attention, or mixture-of-experts models Xu et al. (2023). Although effective, these approaches may underutilize informative representations from intermediate encoder depths. Earlier layers often preserve local and detailed patterns, whereas deeper layers capture more abstract and task-relevant information Lee et al. (2016). In addition, evidence obtained directly from individual modalities may differ from evidence emerging through cross-modal interactions Cheng et al. (2026); Rajendran et al. (2024). Fixed fusion strategies may therefore overlook

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

useful representations or overemphasize redundant and less informative evidence. These limitations give rise to two challenges (Fig. 1).

Challenge 1: Is the most informative representation always the same for each modality? The final encoder output is not necessarily the most informative representation Ukita (2020). Local ECG waveform patterns may remain clearer at early or intermediate depths, whereas useful EHR or CXR information may emerge at deeper stages Smital et al. (2020); Miotto et al. (2016); Qi et al. (2020). The optimal representation can therefore vary across modalities and patients. However, existing approaches often use only the final output or combine all intermediate representations without explicit selection Cui et al. (2023). The former may discard useful information, while the latter may introduce redundancy and noise. An effective model should instead select an informative representation level for each modality and input.

Challenge 2: Is the importance of each modality always the same? The relative importance of EHR, ECG, and CXR may vary across patients Sadeghi et al. (2026); Golovanevsky et al. (2025). CXR findings may provide the primary evidence for one patient, whereas laboratory measurements or ECG patterns may be more informative for another Sen & Bhattacharya (2026). The relevance of cross-modal interaction evidence may also vary across patients Deng et al. (2024). Simple concatenation and fixed weighting cannot adequately reflect these differences. A model should therefore adapt both modality-specific and interaction-specific contributions after selecting informative representations.

To address these challenges, we propose **TriHAR**, a hierarchical adaptive routing framework for heterogeneous clinical data. TriHAR performs three complementary routing operations: Representation-Level Routing selects an informative encoder depth for each modality, Modality-Expert Routing adaptively combines modality-specific predictions, and Interaction-Expert Routing aggregates EHR-centered cross-modal interaction predictions. The resulting branches are subsequently integrated to produce the final clinical prediction.

The main contributions of this work are summarized as follows:

- We propose TriHAR, a hierarchical adaptive routing framework that jointly models representation depth, modality-specific evidence, and cross-modal interaction evidence.
- Representation-Level Routing is introduced to select an informative encoder depth within each modality rather than relying solely on the final encoder output.
- Modality-specific and EHR-centered interaction predictions are formulated as separate expert sets and adaptively combined through Modality-Expert Routing and Interaction-Expert Routing.
- Extensive evaluations across diverse combinations of EHR, ECG, and CXR encoders demonstrate the effectiveness and consistency of TriHAR across heterogeneous representation spaces.

2 Related Work

2.1 Multimodal Learning in the Medical Domain

Multimodal learning has shown considerable potential across clinical tasks, including disease diagnosis, readmission prediction, and mortality prediction Acosta et al. (2022); Wang et al. (2024), by integrating complementary information from heterogeneous data sources to improve predictive performance Wang et al. (2022). Representative fusion methods, including DAFT, MMTM, MedFuse, and HyperFusion, exploit cross-modal complementarity through different fusion mechanisms Wolf et al. (2022); Joze et al. (2020); Hayat et al. (2022); Duenias et al. (2025).

However, many of these methods implicitly treat modality contributions as fixed or comparable rather than explicitly modeling their individual contributions to the final prediction Zhang et al. (2026); Soladoye et al. (2025), even though the reliability and informativeness of each modality may vary across patients and clinical scenarios Jin et al. (2022). This motivates adaptive fusion methods that dynamically estimate and reflect modality-specific contributions.

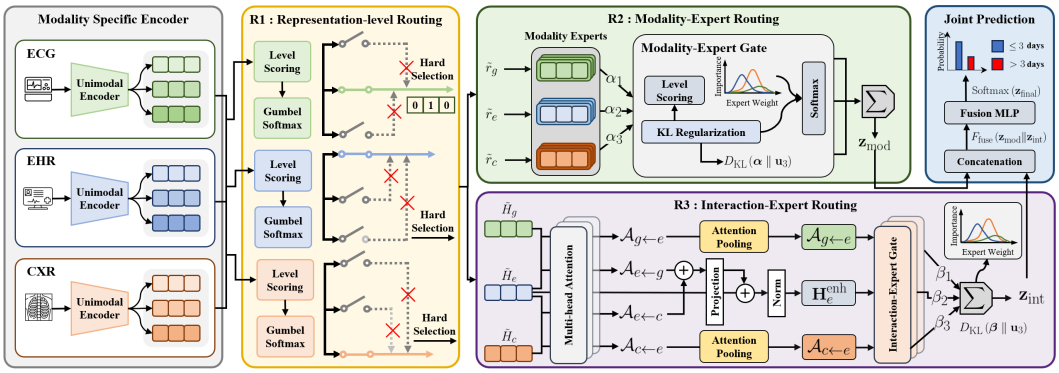


Figure 2: Overall architecture of TriHAR. The framework hierarchically performs Representation-Level Routing (R1), Modality-Expert Routing (R2), and Interaction-Expert Routing (R3). The predictions generated through the R1+R2 and R1+R3 branches are subsequently integrated to produce the final R1+R2 + R3 prediction.

2.2 Mixture-of-Experts

Mixture-of-Experts (MoE) models enable specialized learning by decomposing a network into multiple expert modules Masoudnia & Ebrahimpour (2014); Zhang et al. (2025), with a routing network dynamically assigning or combining experts according to their relevance to each input Jordan & Jacobs (1994). Recent studies have incorporated MoE into multimodal learning to support adaptive fusion Mu & Lin (2025); representative methods, such as I^{2}MoE and FuseMoE, aggregate multimodal representations by estimating modality importance through expert routing Xin et al. (2025); Han et al. (2024).

Nevertheless, these methods generally perform expert routing only after extracting the final high-level representations from modality-specific encoders, overlooking that features from different encoder depths carry distinct semantic information and may contribute differently to downstream prediction Cai et al. (2026). This motivates representation-level routing that explicitly considers multi-level features within each modality rather than relying solely on the final encoder representation.

3 TriHAR

We propose **TriHAR**, a hierarchical adaptive routing framework for integrating Electronic Health Records (EHR), electrocardiograms (ECG), and chest radiographs (CXR). TriHAR accounts for patient-specific variations in informative representation depths, modality contributions, and cross-modal interactions. As illustrated in Fig. 2, it performs three consecutive routing processes. Representation routing selects an encoder level within each modality, while modality and interaction routing adaptively combine direct and relational evidence through two independent soft Mixture-of-Experts (MoE) modules.

Let $\mathcal{M} = \{e, g, c\}$ denote the modality set, where e, g, and c correspond to EHR, ECG, and CXR, respectively.

3.1 Representation-Level Routing

For each modality $m \in \mathcal{M}$, the encoder E_m extracts representations from K depths. We use $K = 3$, corresponding to low-, middle-, and high-level representations. The outputs at level k are defined as follows:

$$\left\{\mathbf{H}_m^{(k)}, \mathbf{r}_m^{(k)}, \mathbf{z}_m^{(k)}\right\}_{k=1}^{K} = E_m\left(\mathbf{x}_m\right), \tag{1}$$

where $\mathbf{x}_m$ denotes the modality input, $\mathbf{H}_m^{(k)}$ denotes the sequence representation, $\mathbf{r}_m^{(k)}$ denotes its routing descriptor, and $\mathbf{z}_m^{(k)}$ denotes the corresponding prediction logits. The sequence representation preserves temporal or spatial information for subsequent cross-modal interaction, whereas the routing descriptor summarizes the representation level for routing.

Since encoder depths may produce different sequence lengths and feature dimensions, the aligned representations are defined as follows:

$$\overline{\mathbf{H}}_m^{(k)} = \mathbf{P}_{m,k}^H \mathcal{R}_L\left(\mathbf{H}_m^{(k)}\right), \qquad \overline{\mathbf{r}}_m^{(k)} = \mathbf{P}_{m,k}^r\left(\mathbf{r}_m^{(k)}\right), \tag{2}$$

where $\mathcal{R}_L(\cdot)$ adjusts the sequence length to L, and $\mathbf{P}_{m,k}^H$ and $\mathbf{P}_{m,k}^r$ project the outputs into a common latent space.

The relevance score of level k and the corresponding routing probabilities are defined as follows:

$$s_m^{(k)} = g_{m,k}\left(\mathbf{r}_m^{(k)}\right), \quad \boldsymbol{\pi}_m = \mathrm{GS}_\tau\left(\mathbf{s}_m\right). \tag{3}$$

where $g_{m,k}(\cdot)$ denotes a level-specific scoring function, $\mathbf{s}_m = [s_m^{(1)}, \ldots, s_m^{(K)}]$ denotes the level-score vector, and $\mathrm{GS}_\tau(\cdot)$ denotes the Gumbel–Softmax operation with temperature τ. A one-hot routing vector $\widehat{\boldsymbol{\pi}}_m$ is obtained via argmax over $\boldsymbol{\pi}_m$ in the forward pass, while gradients are propagated through the soft routing probabilities using the straight-through estimator; this prevents indiscriminate mixing of representation depths while preserving end-to-end optimization. If the routing descriptor retains a sequence dimension, it is averaged before scoring. The selected outputs are defined as follows:

$$\left(\widetilde{\mathbf{H}}_m, \widetilde{\mathbf{r}}_m, \widetilde{\mathbf{z}}_m\right) = \sum_{k=1}^{K} \widehat{\pi}_m^{(k)}\left(\overline{\mathbf{H}}_m^{(k)}, \overline{\mathbf{r}}_m^{(k)}, \mathbf{z}_m^{(k)}\right), \tag{4}$$

where $\widehat{\pi}_m^{(k)}$ denotes the routing indicator for level k. Applying the same decision to all three outputs ensures that subsequent interaction, gating, and prediction operations use information from the same encoder depth.

Representation routing is independently performed for each patient and modality. It requires neither level annotations nor a separate routing loss and is optimized through downstream prediction objectives. TriHAR can therefore select different representation depths according to their patient-specific predictive contributions, directly addressing **Challenge 1**.

3.2 Adaptive Modality and Interaction Routing

After representation routing, TriHAR separately aggregates direct modality evidence and cross-modal interaction evidence. The modality MoE combines predictions derived from the selected modality-specific representations, whereas the interaction MoE combines evidence obtained through EHR-centered cross-modal interactions. The two routing mechanisms are independently parameterized and operate on distinct inputs, but share a common expert-gating formulation. Given a set of expert inputs $\{\mathbf{x}_m\}_{m\in\mathcal{M}}$, a gating function $f(\cdot)$, and expert outputs $\{\mathbf{o}_m\}_{m\in\mathcal{M}}$, the routing score, patient-specific weight, and combined prediction are defined as follows:

$$u_m = f(\mathbf{x}_m), \qquad w_m = \frac{\exp(u_m)}{\sum_{j\in\mathcal{M}}\exp(u_j)}, \qquad \mathbf{z} = \sum_{m\in\mathcal{M}} w_m \mathbf{o}_m. \tag{5}$$

Modality-expert routing and interaction-expert routing instantiate Eq. (5) with different gating inputs, gating functions, and expert outputs, as described below.

3.2.1 Modality-Expert Routing

The selected predictions $\{\widetilde{\mathbf{z}}_m\}_{m\in\mathcal{M}}$ are treated as modality-specific experts. Applying Eq. (5) with gating input $\mathbf{x}_m = \widetilde{\mathbf{r}}_m$, gating function $f_{\mathrm{mod}}(\cdot)$, and expert output $\mathbf{o}_m = \widetilde{\mathbf{z}}_m$ yields the modality-expert weight α_m and the modality-level prediction $\mathbf{z}_{\mathrm{mod}} = \sum_{m\in\mathcal{M}} \alpha_m \widetilde{\mathbf{z}}_m$.

The distribution $\boldsymbol{\alpha} = [\alpha_e, \alpha_g, \alpha_c]$ dynamically determines the relative contributions of EHR, ECG, and CXR for each patient. Because the gate uses representations obtained after representation routing, TriHAR first selects informative evidence within each modality and then determines the contribution of the resulting modality expert.

3.2.2 Interaction-Expert Routing

Direct modality predictions may not fully capture evidence arising from relationships among modalities. TriHAR therefore constructs an interaction MoE using EHR as the central clinical context. For target modality a and source modality b, the output of attention head h is defined as follows:

$$\mathbf{A}_{a\leftarrow b}^{(h)} = \mathrm{Softmax}\left(\frac{\left(\widetilde{\mathbf{H}}_a \mathbf{W}_h^Q\right)\left(\widetilde{\mathbf{H}}_b \mathbf{W}_h^K\right)^\top}{\sqrt{d_h}}\right)\left(\widetilde{\mathbf{H}}_b \mathbf{W}_h^V\right), \tag{6}$$

where $\widetilde{\mathbf{H}}_a$ provides the query, $\widetilde{\mathbf{H}}_b$ provides the key and value, $\mathbf{W}_h^Q$, $\mathbf{W}_h^K$, and $\mathbf{W}_h^V$ denote the corresponding projection matrices, and d_h denotes the head dimension.

The directional interaction representation is defined as follows:

$$\mathcal{A}_{a\leftarrow b} = \mathrm{Concat}\left(\mathbf{A}_{a\leftarrow b}^{(1)}, \ldots, \mathbf{A}_{a\leftarrow b}^{(H)}\right)\mathbf{W}^O, \tag{7}$$

where H denotes the number of attention heads, $\mathbf{W}^O$ denotes the output projection, and $\mathcal{A}_{a\leftarrow b}$ represents target modality a conditioned on source modality b.

TriHAR computes four directional interactions: $\mathcal{A}_{e\leftarrow g}$, $\mathcal{A}_{g\leftarrow e}$, $\mathcal{A}_{e\leftarrow c}$, and $\mathcal{A}_{c\leftarrow e}$. The two EHR-targeted interactions are integrated with the selected EHR representation as follows:

$$\mathbf{H}_e^{\mathrm{enh}} = \mathrm{LN}\left[\widetilde{\mathbf{H}}_e + F_e\left(\mathcal{A}_{e\leftarrow g} \| \mathcal{A}_{e\leftarrow c}\right)\right], \tag{8}$$

where $F_e(\cdot)$ projects the concatenated interactions into the EHR latent space, $\|$ denotes concatenation, and $\mathrm{LN}(\cdot)$ denotes layer normalization. The residual structure preserves the selected EHR information while incorporating complementary ECG and CXR evidence.

The interaction sources are defined as $\mathbf{S}_e = \mathbf{H}_e^{\mathrm{enh}}$, $\mathbf{S}_g = \mathcal{A}_{g\leftarrow e}$, and $\mathbf{S}_c = \mathcal{A}_{c\leftarrow e}$, representing enhanced EHR, EHR-conditioned ECG, and EHR-conditioned CXR, respectively. Each source is converted into an expert representation as follows:

$$a_{m,t} = \frac{\exp\left(\phi_m\left(\mathbf{S}_{m,t}\right)\right)}{\sum_{\ell=1}^{L}\exp\left(\phi_m\left(\mathbf{S}_{m,\ell}\right)\right)}, \qquad \mathbf{q}_m = \sum_{t=1}^{L} a_{m,t}\mathbf{S}_{m,t}, \tag{9}$$

where $\phi_m(\cdot)$ denotes a positional scoring function, $a_{m,t}$ denotes the normalized importance of position t, and $\mathbf{q}_m$ denotes the aggregated representation of interaction expert m. This operation emphasizes informative positions within each interaction source before expert routing.

Applying Eq. (5) to the aggregated expert representations $\{\mathbf{q}_m\}_{m\in\mathcal{M}}$, with gating function $f_{\mathrm{int}}(\cdot)$ and expert output $\mathbf{o}_m = C_m(\mathbf{q}_m)$ (the classifier associated with interaction expert m), yields the interaction-expert weight β_m and the interaction-level prediction $\mathbf{z}_{\mathrm{int}} = \sum_{m\in\mathcal{M}} \beta_m C_m(\mathbf{q}_m)$.

The distribution $\boldsymbol{\beta} = [\beta_e, \beta_g, \beta_c]$ controls the contributions of enhanced EHR, EHR-conditioned ECG, and EHR-conditioned CXR. Thus, modality routing controls direct modality evidence, whereas interaction routing controls cross-modal interaction evidence, allowing TriHAR to adapt both pathways to each patient and thereby address **Challenge 2**.

3.3 Joint Prediction and Optimization

TriHAR maintains the modality and interaction predictions as distinct paths until the final stage. The final prediction is defined as follows:

$$\mathbf{z}_{\text{final}} = F_{\text{fuse}}\left(\mathbf{z}_{\text{mod}} \| \mathbf{z}_{\text{int}}\right), \qquad \widehat{\mathbf{y}} = \text{Softmax}\left(\mathbf{z}_{\text{final}}\right), \tag{10}$$

where $F_{\text{fuse}}(\cdot)$ denotes the learnable fusion function, $\mathbf{z}_{\text{final}}$ denotes the final logits, and $\widehat{\mathbf{y}}$ denotes the predicted class probabilities.

The final, modality, and interaction losses are respectively defined as $\mathcal{L}_{\text{final}} = \text{CE}(\mathbf{z}_{\text{final}}, \mathbf{y})$, $\mathcal{L}_{\text{mod}} = \text{CE}(\mathbf{z}_{\text{mod}}, \mathbf{y})$, and $\mathcal{L}_{\text{int}} = \text{CE}(\mathbf{z}_{\text{int}}, \mathbf{y})$, where $\mathbf{y}$ denotes the ground-truth label. The auxiliary losses encourage both paths to remain independently discriminative.

To prevent excessive concentration on a single expert, the routing regularizers are defined as follows:

$$\mathcal{L}_{\text{mod}}^{\text{route}} = D_{\text{KL}}\left(\boldsymbol{\alpha} \parallel \mathbf{u}_3\right), \qquad \mathcal{L}_{\text{int}}^{\text{route}} = D_{\text{KL}}\left(\boldsymbol{\beta} \parallel \mathbf{u}_3\right), \tag{11}$$

where $\mathbf{u}_3 = [1/3, 1/3, 1/3]$ denotes the uniform expert prior, and these regularizers allow patient-specific expert contributions while discouraging collapse.

The complete objective is defined as follows:

$$\begin{aligned} \mathcal{L}_{\text{TriHAR}} = \mathcal{L}_{\text{final}} + \lambda_{\text{mod}}\mathcal{L}_{\text{mod}} + \lambda_{\text{int}}\mathcal{L}_{\text{int}} \\ + \lambda_{\text{gate}}\left(\mathcal{L}_{\text{mod}}^{\text{route}} + \mathcal{L}_{\text{int}}^{\text{route}}\right), \end{aligned} \tag{12}$$

where λ_{mod} and λ_{int} control the auxiliary prediction losses, and λ_{gate} controls the routing regularization. Representation routing is optimized through the three prediction losses using the straight-through estimator, while modality and interaction routing further receive path-specific supervision and routing regularization, jointly learning patient-specific routing over representation depth, modality evidence, and cross-modal interaction.

4 Experiment

4.1 Experimental Setting

We constructed a trimodal clinical cohort by linking MIMIC-IV, MIMIC-IV-ECG, and MIMIC-CXR-JPG, which provide EHR, ECG, and CXR data, respectively Johnson et al. (2023); Gow et al. (2023); Johnson et al. (2019). ICU length-of-stay prediction was formulated as a binary classification task, with stays of 3 days or less labeled as *short-stay* and longer stays as *long-stay*, following previous MIMIC-based studies Susetzky et al. (2025). All three modalities were temporally aligned with ICU admission, restricted to adult patients' first ICU stay, and preprocessed independently for each encoder input. The resulting cohort comprised 7,504 samples with complete EHR, ECG, and CXR data, divided into training, validation, and test sets at a patient-level ratio of 6:2:2 to prevent data leakage. Full inclusion criteria and modality-specific preprocessing details are provided in Appendix A.

All experiments were implemented in PyTorch and conducted on an NVIDIA RTX PRO 6000 Blackwell GPU, using the AdamW optimizer with early stopping based on validation AUROC (patience of 10 epochs). Differentiable hard representation routing was implemented via Straight-Through Gumbel–Softmax, with KL-divergence regularization toward a uniform distribution applied to both routing distributions. Each experiment was repeated using three random seeds, and performance was evaluated using Accuracy, AUROC, and Precision. Full architectural and optimization hyperparameters, including batch size, latent dimension, learning rate, weight decay, temperature, and regularization weights, are provided in Appendix C (Table 5).

For fair comparison, all baselines used the same modality-specific encoders, preprocessing, patient-level splits, training protocol, and evaluation criteria as TriHAR. We compared TriHAR with eight multimodal learning methods: I^2MoE, FuseMoE, DDMF, and MulT in the trimodal setting Xin et al. (2025); Han et al. (2024); Nan et al. (2026); Tsai et al. (2019), and DAFT, MMTM, MedFuse, and HyperFusion in the bimodal settings Wolf et al. (2022); Joze et al. (2020); Hayat et al. (2022); Duenias et al. (2025).

Table 1: Performance comparison of trimodal fusion methods across 27 combinations of EHR, ECG, and CXR encoders. Best and second-best results per configuration are highlighted in light green and light yellow, respectively, with boldface and underlining.

EHR	ECG	CXR	I²MoE Xin et al. (2025)			FuseMoE Han et al. (2024)			DDMF Nan et al. (2026)			MulT Tsai et al. (2019)			TriHAR (ours)		
			ACC	AUROC	PREC	ACC	AUROC	PREC	ACC	AUROC	PREC	ACC	AUROC	PREC	ACC	AUROC	PREC
		ResNet18	0.632	0.646	0.606	0.633	0.652	0.610	0.616	0.636	0.595	**0.642**	**0.660**	**0.618**	0.634	0.657	0.610
	WaveNet	DenseNet	0.625	0.648	0.606	0.640	0.658	0.615	0.640	**0.669**	0.616	0.630	0.660	**0.616**	**0.642**	0.656	0.616
		VGGNet	0.652	0.682	0.631	0.628	0.652	0.608	0.639	0.659	0.617	0.623	0.651	0.607	**0.657**	**0.682**	**0.643**
		ResNet18	0.638	0.660	**0.614**	0.636	**0.663**	0.613	0.633	0.653	0.613	0.618	0.654	0.606	**0.638**	0.655	0.612
FTTrans	1D-ResNet	DenseNet	0.632	0.651	0.611	0.634	0.657	0.609	**0.638**	**0.659**	**0.616**	0.625	0.654	0.602	0.636	0.652	0.607
		VGGNet	0.620	0.644	0.606	0.644	0.661	0.618	0.634	0.668	0.624	0.623	0.652	0.605	**0.645**	**0.670**	**0.626**
		ResNet18	0.615	0.651	0.600	**0.630**	0.655	**0.609**	0.625	**0.657**	0.598	0.630	0.656	0.605	0.625	0.645	0.602
	LSTM	DenseNet	0.647	0.671	0.627	0.644	0.661	0.618	**0.649**	**0.677**	**0.631**	0.633	0.665	0.617	0.627	0.655	0.610
		VGGNet	0.643	0.663	0.618	0.643	0.657	0.618	0.638	0.651	0.614	0.636	0.669	0.619	**0.656**	**0.674**	**0.634**
		ResNet18	0.626	**0.646**	**0.603**	0.611	0.634	0.597	0.603	0.627	0.588	0.600	0.644	0.597	**0.633**	0.642	0.603
	WaveNet	DenseNet	**0.646**	**0.664**	**0.623**	0.638	0.660	0.612	0.589	0.616	0.582	0.634	0.660	0.610	0.638	0.642	0.611
		VGGNet	0.630	0.648	0.605	0.632	0.641	0.606	0.592	0.615	0.579	0.632	0.655	0.611	**0.650**	**0.683**	**0.633**
		ResNet18	**0.642**	**0.666**	**0.616**	0.626	0.646	0.608	0.625	0.624	0.596	0.636	0.658	0.609	0.614	0.642	0.601
TabTrans	1D-ResNet	DenseNet	0.616	0.631	0.594	0.625	0.649	0.601	0.618	0.630	0.593	**0.642**	**0.661**	**0.621**	0.629	0.642	0.620
		VGGNet	0.618	0.636	0.600	0.618	0.639	0.588	0.632	0.638	0.609	0.605	0.651	0.608	**0.658**	**0.668**	**0.636**
		ResNet18	0.610	0.623	0.590	0.623	0.642	0.600	0.609	0.627	0.592	0.618	0.624	0.594	**0.632**	**0.647**	**0.608**
	LSTM	DenseNet	0.626	0.645	0.605	0.607	0.621	0.590	0.594	0.614	0.575	**0.630**	**0.652**	0.606	0.627	0.629	**0.608**
		VGGNet	0.634	0.651	0.613	0.624	0.632	0.595	0.627	0.645	0.602	0.638	0.654	0.613	**0.642**	**0.670**	**0.637**
		ResNet18	0.630	0.647	0.601	0.626	0.636	0.603	0.613	0.633	0.594	0.605	0.619	0.577	**0.639**	**0.656**	**0.617**
	WaveNet	DenseNet	0.625	0.643	0.603	0.618	0.640	0.598	0.618	0.628	0.590	0.618	**0.645**	0.600	**0.633**	0.639	**0.611**
		VGGNet	0.615	0.647	0.598	0.604	0.605	0.577	0.642	0.662	0.616	**0.651**	**0.678**	**0.631**	0.639	0.663	0.621
		ResNet18	**0.631**	**0.660**	**0.621**	0.619	0.640	0.599	0.601	0.619	0.584	0.595	0.633	0.586	0.611	0.639	0.591
TPC	1D-ResNet	DenseNet	0.621	**0.651**	**0.605**	**0.634**	0.649	0.604	0.615	0.627	0.593	0.622	0.648	0.599	0.630	0.644	0.603
		VGGNet	0.626	0.643	0.604	0.620	0.632	0.593	0.616	0.648	0.599	0.642	0.666	0.619	**0.644**	**0.677**	**0.625**
		ResNet18	0.625	**0.652**	0.611	0.608	0.633	0.590	0.628	0.645	0.602	**0.630**	0.648	0.602	0.619	0.650	**0.651**
	LSTM	DenseNet	**0.634**	**0.660**	**0.614**	0.610	0.630	0.595	0.628	0.654	0.612	0.628	0.650	0.608	0.619	0.655	0.503
		VGGNet	0.622	0.648	0.605	0.621	0.646	0.598	0.641	0.665	0.627	**0.643**	0.657	0.618	0.627	**0.676**	**0.678**
Average Rank ↓			3.037	2.852	2.741	3.222	3.407	3.296	3.519	3.667	3.741	3.148	2.593	3.037	**2.074**	**2.481**	**2.185**

4.2 Experimental Results

4.2.1 Trimodal Performance Comparison

Table 1 compares TriHAR with four trimodal fusion baselines across 27 combinations of EHR, ECG, and CXR encoders (full per-configuration distribution in Appendix Figure 5). No single baseline is uniformly best; the top-performing method changes from one encoder combination to another, reflecting the sensitivity of existing fusion methods to encoder choices. TriHAR instead maintains the best overall average rank for all three metrics (2.074, 2.481, and 2.185 for Accuracy, AUROC, and Precision, respectively), indicating that adaptively selecting representation depth and reweighting modality contributions allows the model to remain consistently competitive across heterogeneous encoder combinations rather than depending on a specific one.

4.2.2 Bimodal Performance Comparison

To evaluate TriHAR with two available modalities, we considered the EHR–ECG and EHR–CXR settings. For each setting, we retrained TriHAR using only the available modalities; the unavailable modality was zero-masked, and its attention outputs and expert weights were excluded from prediction. As in the trimodal setting, the strongest bimodal baseline varies by configuration and modality pairing (Table 2), whereas TriHAR consistently ranks highly across both settings (best average rank of 2.028, 2.056, and 2.222 for Accuracy, AUROC, and Precision, respectively; full per-configuration distributions in Appendix Figures 6 and 7). This indicates that TriHAR's routing remains competitive when only one auxiliary modality is available, supporting the broader case for hierarchical adaptive routing over fixed fusion in clinical multimodal prediction.

4.2.3 Routing Behavior Analysis

Table 3 summarizes the average representation-level selection ratios for each modality-specific encoder (the full per-configuration heatmap is provided in Appendix Figure 8), showing that selection patterns vary across both modalities and encoder architectures: for EHR, FTTrans and TabTrans favor high-level representations, whereas TPC shows no strong preference among the three levels; for

Table 2: Performance comparison in the EHR–ECG and EHR–CXR settings with one unavailable auxiliary modality. Best and second-best results per configuration are highlighted in light green and light yellow, respectively, with boldface and underlining.

EHR	ECG	CXR	DAFT Wolf et al. (2022)			MMTM Joze et al. (2020)			MedFuse (Hayat et al. 2022)			HyperFusion Duenias et al. (2025)			TriHAR (ours)		
			ACC	AUROC	PREC	ACC	AUROC	PREC	ACC	AUROC	PREC	ACC	AUROC	PREC	ACC	AUROC	PREC
FTTrans	WaveNet	N/A	**0.624**	0.625	0.594	0.622	**0.627**	0.591	0.623	0.618	**0.595**	0.600	0.612	0.573	0.621	0.622	0.589
	1D-ResNet	N/A	0.618	0.625	0.587	**0.629**	0.623	**0.598**	0.624	**0.628**	0.596	0.600	0.610	0.580	0.622	0.623	0.597
	LSTM	N/A	0.615	0.629	0.587	0.624	0.628	0.592	0.625	0.630	0.592	0.620	**0.636**	0.595	**0.630**	0.627	**0.600**
	N/A	ResNet18	0.631	0.647	0.607	**0.638**	0.652	0.613	0.633	**0.658**	0.614	0.632	0.641	0.604	0.631	0.657	**0.627**
	N/A	DenseNet	0.638	0.658	0.616	**0.642**	0.666	**0.621**	0.636	**0.666**	0.617	0.625	0.649	0.604	0.642	0.658	0.618
	N/A	VGGNet	0.647	0.678	0.632	0.644	0.674	0.626	0.654	0.683	0.635	0.657	0.671	0.635	**0.662**	**0.687**	**0.643**
TabTrans	WaveNet	N/A	0.550	0.536	0.528	0.545	0.536	0.527	0.537	0.544	0.524	0.581	0.577	**0.554**	**0.612**	**0.589**	0.490
	1D-ResNet	N/A	0.563	0.550	0.535	0.576	0.561	0.543	0.548	0.532	0.527	0.598	0.574	**0.560**	**0.609**	**0.592**	0.475
	LSTM	N/A	0.553	0.546	0.531	0.567	0.553	0.534	0.548	0.545	0.524	0.591	0.554	0.552	**0.612**	**0.578**	**0.672**
	N/A	ResNet18	0.604	0.620	0.583	0.607	0.615	0.585	0.616	0.625	0.594	0.597	0.642	0.600	**0.642**	**0.647**	**0.616**
	N/A	DenseNet	0.610	0.632	0.590	0.623	0.641	0.604	**0.628**	0.645	0.606	0.591	0.614	0.576	**0.628**	**0.652**	**0.608**
	N/A	VGGNet	0.640	0.670	0.624	0.644	0.660	0.623	**0.647**	0.669	0.623	0.644	0.663	0.618	0.644	**0.671**	**0.636**
TPC	WaveNet	N/A	0.612	0.576	0.571	0.580	0.569	0.549	0.581	0.582	0.556	0.571	0.561	0.543	**0.618**	**0.608**	**0.588**
	1D-ResNet	N/A	0.580	0.574	0.553	0.606	**0.584**	**0.573**	0.580	0.571	0.544	0.576	0.572	0.549	**0.615**	0.568	0.553
	LSTM	N/A	0.608	0.579	0.566	0.612	0.602	0.577	0.596	0.575	0.563	0.588	0.551	0.547	**0.616**	**0.606**	**0.593**
	N/A	ResNet18	**0.626**	**0.652**	0.600	0.613	0.642	0.597	0.607	0.629	0.593	0.604	0.615	0.587	0.615	0.646	**0.616**
	N/A	DenseNet	0.629	0.648	0.601	0.627	0.641	0.599	**0.645**	**0.673**	**0.621**	0.611	0.624	0.588	0.613	0.643	0.598
	N/A	VGGNet	0.637	0.670	**0.623**	0.629	0.665	0.621	**0.641**	0.656	0.619	0.633	0.668	0.621	0.579	**0.674**	0.618
Average Rank ↓			3.278	2.944	3.056	2.889	3.278	2.778	2.861	2.889	3.167	3.944	3.833	3.778	**2.028**	**2.056**	**2.222**

ECG, WaveNet and 1D-ResNet most frequently select low-level representations, while LSTM favors high-level representations; and all CXR encoders predominantly select high-level representations, particularly ResNet18, with a selection ratio exceeding 90%.

Table 4 presents the average expert utilization patterns of Modality-Expert Routing (R2) and Interaction-Expert Routing (R3): R2 adaptively combines direct predictions from the selected EHR, ECG, and CXR representations, whereas R3 combines predictions from enhanced EHR, EHR-conditioned ECG, and EHR-conditioned CXR, with the enhanced EHR expert incorporating complementary ECG and CXR information; the full per-configuration heatmap is provided in Appendix Figure 9.

R2 assigns its largest average weight to CXR at 58.80%, whereas R3 assigns its largest average weight to enhanced EHR at 39.44%. These distinct utilization patterns indicate that R2 and R3 emphasize different experts when routing direct modality evidence and cross-modal interaction evidence, respectively, supporting the separate adaptive routing pathways introduced to address **Challenge 2**.

Table 3: Average representation-level selection ratios for each modality-specific encoder.

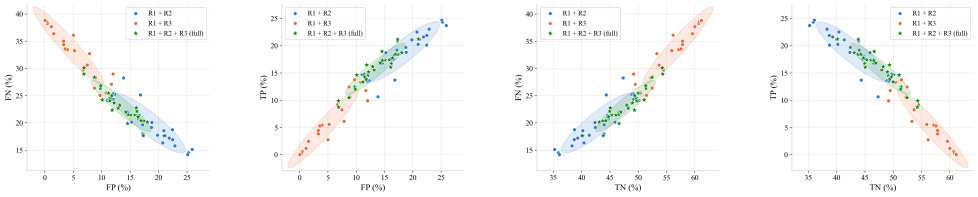
Modality	Encoder	Low (%)	Mid (%)	High (%)
EHR	FTTrans	30.37 ± 5.97	29.02 ± 1.95	$\mathbf{40.62 \pm 4.03}$
	TabTrans	35.74 ± 8.21	21.94 ± 4.76	$\mathbf{42.32 \pm 4.69}$
	TPC	34.82 ± 12.35	30.18 ± 11.71	$\mathbf{35.00 \pm 2.90}$
ECG	WaveNet	$\mathbf{35.33 \pm 4.54}$	31.89 ± 9.30	32.77 ± 11.99
	1D-ResNet	$\mathbf{39.49 \pm 2.41}$	35.18 ± 2.56	25.34 ± 3.32
	LSTM	23.05 ± 9.30	37.95 ± 12.29	$\mathbf{39.00 \pm 4.32}$
CXR	ResNet18	3.55 ± 3.29	6.20 ± 6.49	$\mathbf{90.24 \pm 9.73}$
	DenseNet	5.60 ± 6.01	6.72 ± 7.54	$\mathbf{87.68 \pm 13.56}$
	VGGNet	0.30 ± 0.26	40.91 ± 7.90	$\mathbf{58.79 \pm 7.65}$

Table 4: Average expert utilization under R1+R2 and R1+R3. In R1+R3, EHR denotes the enhanced representation incorporating ECG and CXR information.

Routing Config.	EHR (%)	ECG (%)	CXR (%)
R1+R2	22.67 ± 2.99	18.54 ± 4.14	$\mathbf{58.80 \pm 6.63}$
R1+R3	$\mathbf{39.44 \pm 2.75}$	30.43 ± 1.23	30.13 ± 1.53

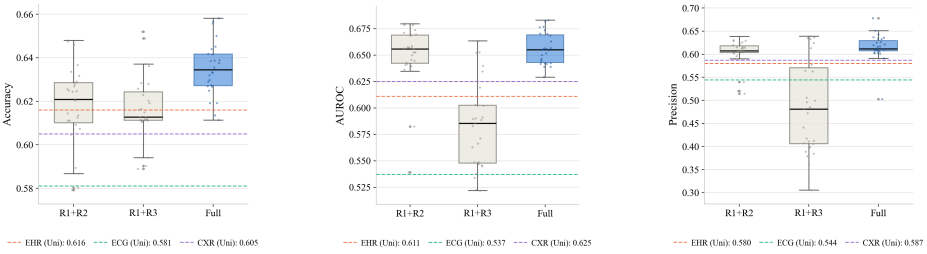
4.2.4 Component and Prediction Analysis

Figure 3 compares the confusion-matrix composition of R1+R2, R1+R3, and the full model across all 27 encoder combinations, as a percentage of the fixed $N = 1501$ test set (the encoder-level breakdown is provided in Appendix Figure 10). R1+R2 has the highest mean TP proportion (18.8%) and the lowest mean FN proportion (20.0%) of the test set, indicating greater sensitivity to long-stay patients, whereas R1+R3 has the highest mean TN proportion (57.2%) and the lowest mean FP proportion (4.0%) of the test set, favoring short-stay identification. The full model falls between both extremes on every component (TN 48.5%, FP 12.6%, FN 23.9%, TP 15.0%), confirming that R1+R2 and R1+R3 are complementary rather than redundant and supporting their design as separate adaptive routing pathways introduced to address **Challenge 2**.



(a) (b) (c) (d)

Figure 3: Distribution of confusion-matrix components across all 27 encoder combinations, expressed as a percentage of the fixed test set, for the R1+R2, R1+R3, and full R1+R2+R3 routing configurations (ellipses denote the 1.5 standard-deviation region for each configuration): (a) FP vs. FN, (b) FP vs. TP, (c) TN vs. FN, (d) TN vs. TP.



(a) (b) (c)

Figure 4: Per-configuration distributions of (a) Accuracy, (b) AUROC, and (c) Precision across the 27 encoder combinations, for the R1+R2, R1+R3, and full R1+R2+R3 routing configurations (the full model is highlighted in blue).

Figure 4 compares the per-configuration Accuracy, AUROC, and Precision of R1+R2, R1+R3, and the full model against the three unimodal baselines, pooled across the 27 encoder combinations (the per-configuration view is provided in Appendix Figure 11). R1+R2 exceeds all three baselines on average (Accuracy 0.618, AUROC 0.649, Precision 0.603), while R1+R3 is competitive on Accuracy (0.617) but weaker on AUROC (0.583) and Precision (0.487). By jointly incorporating R2 and R3, the full model achieves the best overall performance (Accuracy 0.635, AUROC 0.656, Precision 0.615), surpassing the strongest unimodal baseline in 25, 27, and 26 of the 27 combinations, respectively, providing direct evidence for the benefit of the full hierarchical adaptive routing framework.

5 Conclusion

In this study, we proposed TriHAR, a hierarchical adaptive routing framework for clinical prediction using EHR, ECG, and CXR data. TriHAR selects informative encoder depths through Representation-Level Routing (R1) and integrates direct modality and EHR-centered cross-modal interaction evidence through Modality-Expert Routing (R2) and Interaction-Expert Routing (R3), respectively. Experiments on ICU length-of-stay prediction showed that TriHAR achieved the best average ranks across diverse encoder combinations and remained effective with one auxiliary modality. Routing analyses further revealed modality- and encoder-dependent representation preferences and complementary prediction characteristics between R2 and R3. These findings demonstrate the effectiveness of jointly adapting representation depth, modality contributions, and cross-modal interactions for heterogeneous multimodal clinical prediction. This work is validated on a single trimodal cohort from MIMIC-IV, MIMIC-IV-ECG, and MIMIC-CXR-JPG, and to our knowledge, no EHR-ECG-CXR resource independent of this MIMIC ecosystem currently exists. Separately, the lower performance observed for some encoder combinations may reflect that a given encoder architecture does not integrate well with a particular modality, suggesting that larger or higher-capacity encoders for these modalities could be a promising direction for future improvement.

AI USE STATEMENT

In this work, we used generative AI tools for tasks with required disclosure, specifically assisting with translation between Korean and English during manuscript preparation. We also used generative AI tools for tasks with recommended disclosure, specifically grammar and language polishing to improve readability. We have reviewed all AI-assisted text for accuracy and consistency with our intended meaning. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study uses the MIMIC-IV, MIMIC-IV-ECG, and MIMIC-CXR-JPG databases, which contain de-identified electronic health records, electrocardiograms, and chest radiographs collected under IRB-approved protocols with a waiver of informed consent for retrospective research use. Access to these datasets was obtained after completing the required data use training and certification, and all data were used in accordance with the corresponding data use agreements. As TriHAR is intended for research purposes at this stage and has not been validated for clinical deployment, we caution against its use in real-world clinical decision-making without further prospective validation and fairness auditing across patient subgroups.

REPRODUCIBILITY STATEMENT

To support reproducibility, the complete implementation is provided as anonymized supplementary material and will additionally be released at an anonymous public repository (https://anonymous.4open.science/r/TriHAR) upon publication. Full dataset construction and preprocessing details for MIMIC-IV, MIMIC-IV-ECG, and MIMIC-CXR-JPG are provided in Appendix A, and complete architectural and optimization hyperparameters are provided in Appendix C (Table 5).

REFERENCES

Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. Multimodal biomedical ai. *Nature medicine*, 28(9):1773–1784, 2022.

Odeh Victor Adeyi, Zhang Xiaoling, Kuyikabasi Uko, and Izis Kanjarawy. Multimodal precision diagnostics via cardionet+: fusing time-series, imaging, and knowledge graphs for cardiac risk modeling. *Biomedical Signal Processing and Control*, 126:110916, 2026.

Selcan Kaplan Berkaya, Alper Kursat Uysal, Efnan Sora Gunal, Semih Ergin, Serkan Gunal, and M Bilginer Gulmezoglu. A survey on ecg analysis. *Biomedical Signal Processing and Control*, 43:216–235, 2018.

Qing Cai, Fan Zhang, Cheng Zhang, Zhi Liu, et al. Semc: Structure-enhanced mixture-of-experts contrastive learning for ultrasound standard plane recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2543–2551, 2026.

Erdi Çallı, Ecem Sogancioglu, Bram Van Ginneken, Kicky G Van Leeuwen, and Keelin Murphy. Deep learning for chest x-ray analysis: A survey. *Medical image analysis*, 72:102125, 2021.

Cheng Cheng, Wenzhe Liu, Ruisi Shang, Yikang Cheng, and Ziyu Jia. Modality-aware fusion with alignment and balancing for robust multimodal emotion recognition. *Information Fusion*, pp. 104368, 2026.

Can Cui, Haichun Yang, Yaohong Wang, Shilin Zhao, Zuhayr Asad, Lori A Coburn, Keith T Wilson, Bennett A Landman, and Yuankai Huo. Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review. *Progress in Biomedical Engineering*, 5(2):022001, 2023.

Ruining Deng, Nazim Shaikh, Gareth Shannon, and Yao Nie. Cross-modality attention-based multimodal fusion for non-small cell lung cancer (nsclc) patient survival prediction. In *Medical Imaging 2024: Digital and Computational Pathology*, volume 12933, pp. 46–50. SPIE, 2024.

Daniel Duenias, Brennan Nichyporuk, Tal Arbel, Tammy Riklin Raviv, et al. Hyperfusion: A hypernetwork approach to multimodal integration of tabular and medical imaging data for predictive modeling. *Medical Image Analysis*, 102:103503, 2025.

Michal Golovanevsky, Pranav Mahableshwarkar, Carsten Eickhoff, and Ritambhara Singh. Picme: Pipeline for contrastive modality evaluation and encoding in the mimic dataset. *arXiv preprint arXiv:2507.03165*, 2025.

Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W Waks, Parastou Eslami, Tanner Carbonati, et al. Mimic-iv-ecg: Diagnostic electrocardiogram matched subset. *Type: dataset*, 6:13–14, 2023.

Xing Han, Huy Nguyen, Carl Harris, Nhat Ho, and Suchi Saria. Fusemoe: Mixture-of-experts transformers for fleximodal fusion. *Advances in Neural Information Processing Systems*, 37: 67850–67900, 2024.

Nasir Hayat, Krzysztof J Geras, and Farah E Shamout. Medfuse: Multi-modal fusion with clinical time-series data and chest x-ray images. In *Machine Learning for Healthcare Conference*, pp. 479–503. PMLR, 2022.

Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. Healnet: Multimodal fusion for heterogeneous biomedical data. *Advances in Neural Information Processing Systems*, 37:64479–64498, 2024.

Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. Evaluating explainable ai on a multi-modal medical imaging task: Can existing algorithms fulfill clinical requirements? In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 11945–11953, 2022.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 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.

Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

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*, pp. 13289–13299, 2020.

Gayoung Lee, Yu-Wing Tai, and Junmo Kim. Deep saliency with encoded low level distance map and high level features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 660–668, 2016.

Junwei Liu, Xiaoping Cen, Chenxin Yi, Feng-ao Wang, Junxiang Ding, Jinyu Cheng, Qinhua Wu, Baowen Gai, Yiwen Zhou, Ruikun He, et al. Challenges in ai-driven biomedical multimodal data fusion and analysis. *Genomics, proteomics & bioinformatics*, 23(1):qzaf011, 2025.

Saeed Masoudnia and Reza Ebrahimpour. Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42(2):275–293, 2014.

Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1): 26094, 2016.

Siyuan Mu and Sen Lin. A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications. *arXiv preprint arXiv:2503.07137*, 2025.

Ke Nan, Maggie Samaan, Benjamin Burns, Xia Ning, Yuchi Han, and Yuan Xue. Dual-domain multimodal hyperbolic fusion for cardiopulmonary disease diagnosis in emergency care. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 8700–8710, 2026.

Xiao Qi, Lloyd Brown, David J Foran, and Ilker Hacihaliloglu. Chest x-ray image phase features for improved diagnosis of covid-19 using convolutional neural network. *arXiv preprint arXiv:2011.03585*, 2020.

Suraj Rajendran, Weishen Pan, Mert R Sabuncu, Yong Chen, Jiayu Zhou, and FEI Wang. Learning across diverse biomedical data modalities and cohorts: Challenges and opportunities for innovation. *Patterns*, 5(2), 2024.

Ali Rasekh, Reza Heidari, Amir Hosein Haji Mohammad Rezaie, Parsa Sharifi Sedeh, Zahra Ahmadi, Prasenjit Mitra, and Wolfgang Nejdl. Robust fusion of time series and image data for improved multimodal clinical prediction. *IEEE Access*, 12:174107–174121, 2024.

Alireza Sadeghi, Farshid Hajati, Ahmadreza Argha, Nigel H Lovell, Min Yang, and Hamid Alinejad-Rokny. Interpretable graph-based models on multimodal biomedical data integration: A technical review and benchmarking. *Nature Communications*, 2026.

Jayanta Sen and Sweta Bhattacharya. Xai in heart disease: A review of concepts, applications and limitations. *Archives of Computational Methods in Engineering*, pp. 1–35, 2026.

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.

Lukas Smital, Clifton R Haider, Martin Vitek, Pavel Leinveber, Pavel Jurak, Andrea Nemcova, Radovan Smisek, Lucie Marsanova, Ivo Provaznik, Christopher L Felton, et al. Real-time quality assessment of long-term ecg signals recorded by wearables in free-living conditions. *IEEE Transactions on Biomedical Engineering*, 67(10):2721–2734, 2020.

Afeez Adekunle Soladoye, Nicholas Aderinto, Damilola Osho, and David B Olawade. Explainable machine learning models for early alzheimer's disease detection using multimodal clinical data. *International journal of medical informatics*, pp. 106093, 2025.

Tobias Susetzky, Huaqi Qiu, Rickmer Braren, and Daniel Rueckert. A holistic time-aware classification model for multimodal longitudinal patient data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 24–33. Springer, 2025.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 6558–6569, 2019.

Jumpei Ukita. Causal importance of low-level feature selectivity for generalization in image recognition. *Neural Networks*, 125:185–193, 2020.

Chutong Wang, Xuebing Yang, Mengxuan Sun, Yifan Gu, Jinghao Niu, and Wensheng Zhang. Multimodal fusion network for icu patient outcome prediction. *Neural Networks*, 180:106672, 2024.

Daheng Wang, Tong Zhao, Wenhao Yu, Nitesh V Chawla, and Meng Jiang. Deep multimodal complementarity learning. *IEEE Transactions on Neural Networks and Learning Systems*, 34 (12):10213–10224, 2022.

Tom Nuno Wolf, Sebastian Pölsterl, Christian Wachinger, Alzheimer's Disease Neuroimaging Initiative, et al. Daft: A universal module to interweave tabular data and 3d images in cnns. *NeuroImage*, 260:119505, 2022.

Jiayi Xin, Sukwon Yun, Jie Peng, Inyoung Choi, Jenna L Ballard, Tianlong Chen, and Qi Long. I2moe: Interpretable multimodal interaction-aware mixture-of-experts. *arXiv preprint arXiv:2505.19190*, 2025.

Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132, 2023.

Danyang Zhang, Junhao Song, Ziqian Bi, Xinyuan Song, Yingfang Yuan, Tianyang Wang, Joe Yeong, and Junfeng Hao. Mixture of experts in large language models. *arXiv preprint arXiv:2507.11181*, 2025.

Qingyang Zhang, Yake Wei, Zongbo Han, Huazhu Fu, Xi Peng, Qinghua Hu, Cheng Deng, Cai Xu, Jie Wen, Di Hu, et al. Multimodal fusion on low-quality data: A comprehensive survey. *Information Fusion*, pp. 104437, 2026.

A Additional Details on Dataset construction and preprocessing

A.1 Description of the clinical prediction task

The proposed framework was evaluated on a trimodal clinical cohort constructed by linking electronic health record (EHR), electrocardiogram (ECG), and chest radiograph (CXR) data. These modalities provide complementary descriptions of a patient's clinical condition. EHR data contain structured demographic, physiological, laboratory, and treatment-related information; ECG data capture cardiac electrical activity; and CXR data provide anatomical and pathological information from the thoracic region.

We formulated intensive care unit (ICU) length-of-stay prediction as a binary classification task. Patients with an ICU stay of three days or less were assigned to the short-stay class, whereas patients with an ICU stay longer than three days were assigned to the long-stay class. Accordingly, the target label was defined as:

$$y = \begin{cases} 0, & \text{if ICU length of stay} \leq 3 \text{ days}, \\ 1, & \text{if ICU length of stay} > 3 \text{ days}. \end{cases} \tag{13}$$

The three-day threshold follows adopted settings in previous ICU length-of-stay prediction studies Susetzky et al. (2025). Only adult patients for whom all three modalities were available were included in the primary trimodal cohort.

Each multimodal sample was centered on a single ICU stay and consisted of clinical information collected during the initial period of ICU admission. To reduce potential dependencies among repeated admissions, only the first ICU stay of each patient was retained.

A.2 Datasets

MIMIC-IV Johnson et al. (2023). The EHR modality was obtained from MIMIC-IV version 3.1, a large-scale, publicly available and deidentified clinical database containing comprehensive information on patients admitted to the emergency department and ICU at Beth Israel Deaconess Medical Center. The database includes demographics, laboratory measurements, vital signs, medications, procedures, and clinical outcomes.

We included adult patients aged 18 years or older and retained only the first ICU stay for each patient. Fifteen continuous clinical variables and one categorical variable (gender) were extracted from the first 24 hours after ICU admission, including demographic, physiological, and laboratory information. The 15 continuous clinical variables were age, heart rate, systolic blood pressure, diastolic blood pressure, respiratory rate, temperature, peripheral oxygen saturation (SpO_2), blood urea nitrogen (BUN), creatinine, sodium, potassium, bicarbonate, white blood cell count (WBC), hemoglobin, and anion gap. Only patients with complete observations for all 15 continuous clinical variables and gender were retained. Continuous variables were standardized using statistics computed from the training set and subsequently applied to the validation and test sets to prevent information leakage. Gender was treated as a categorical variable and converted into an integer index using label encoding.

MIMIC-IV-ECG Gow et al. (2023). The ECG modality was obtained from MIMIC-IV-ECG, which contains approximately 800,000 diagnostic 12-lead ECG recordings from nearly 160,000 patients. The dataset provides raw waveform signals together with acquisition time and clinical metadata, enabling direct linkage with the corresponding hospital admissions in MIMIC-IV. Each recording has a duration of 10 seconds and is sampled at 500 Hz.

ECG records were linked to the corresponding EHR records using the hospital admission identifier (hadm_id). When multiple ECG recordings were available for a single admission, the recording acquired closest to the ICU admission time was selected to best reflect the patient's physiological condition near ICU admission. Signals containing missing values were excluded. Recordings shorter than 5,000 time steps were zero-padded, whereas longer recordings were truncated to 5,000 time steps. All 12 leads were retained, resulting in an input waveform of $12 \times 5{,}000$ samples for each

ECG recording. Finally, each ECG channel was standardized using statistics computed from the training set and applied unchanged to the validation and test sets.

MIMIC-CXR-JPG Johnson et al. (2019). The CXR modality was obtained from MIMIC-CXR-JPG, which contains 377,110 chest radiographs converted from the original MIMIC-CXR DICOM images. The dataset contains chest radiographs acquired from a diverse patient population with a wide range of thoracic diseases and clinical conditions, making it a representative benchmark for medical image analysis and multimodal clinical learning.

Only frontal chest radiographs acquired in the anteroposterior (AP) or posteroanterior (PA) view were considered. Images were first matched to MIMIC-IV records at the patient level, and only those acquired within ± 24 hours of ICU admission were retained to ensure temporal consistency across modalities. When multiple eligible radiographs were available, the image acquired closest to the ICU admission time was selected. Each image was converted to RGB format, resized to 224×224 pixels, and normalized using the standard ImageNet mean and standard deviation before being used for model training.

A.3 Trimodal cohort construction

Following the inclusion criteria described in A.2, the final trimodal cohort was constructed by retaining adult patients with complete EHR, ECG, and CXR observations. For each patient, only the first ICU stay was retained. The EHR variables collected during the first 24 hours after ICU admission, the ECG recording closest to ICU admission, and the frontal CXR closest to ICU admission were aligned and combined into a single multimodal sample. Samples for which the corresponding ECG waveform or CXR image could not be located or loaded were excluded. ECG recordings containing missing values were also excluded.

The resulting cohort consisted of 7,504 complete trimodal samples. Each sample corresponded to one patient's first ICU stay and contained the following modalities:

1. structured EHR variables collected during the first 24 hours after ICU admission;
2. a temporally aligned 12-lead ECG recording; and
3. a temporally aligned frontal CXR image.

Based on the three-day ICU length-of-stay threshold defined in A.2, the final cohort included 4,587 (61.1%) short-stay samples (≤ 3 days) and 2,917 (38.9%) long-stay samples (> 3 days).

The cohort was divided into training, validation, and test sets using a stratified 6:2:2 split at the patient level. Stratification was performed according to the ICU length-of-stay class to preserve the class distribution across the three subsets. All preprocessing statistics were estimated exclusively from the training set and subsequently applied unchanged to the validation and test sets to prevent information leakage.

B Detailed description of comparison methods

We compared the proposed framework with representative multimodal fusion and mixture-of-experts methods. To provide a controlled comparison, all methods used the same modality-specific EHR, ECG, and CXR encoders for each encoder configuration. The preprocessing procedures, data splits, training protocol, and evaluation criteria were also kept identical across methods.

B.1 Trimodal fusion baselines

The trimodal comparison included I^2MoE, FuseMoE, DDMF, and MulT Xin et al. (2025); Han et al. (2024); Nan et al. (2026); Tsai et al. (2019). These methods represent different approaches to adaptive expert routing, multimodal interaction modeling, and transformer-based fusion.

- I^2MoE Xin et al. (2025) is an interaction-aware mixture-of-experts framework designed to model both modality-specific information and interactions among modalities. Rather than treating multimodal inputs as a single concatenated representation, I^2MoE constructs

specialized experts for different information sources and uses a routing mechanism to determine their relative contributions.

- FuseMoE Han et al. (2024) is a multimodal mixture-of-experts method that uses learnable routing to assign multimodal inputs to specialized fusion experts. The gating network estimates the relevance of the available experts for each sample, allowing the model to adapt the fusion process to input-specific modality characteristics.
- DDMF Nan et al. (2026) is a deep multimodal fusion framework that learns interactions among heterogeneous modality representations through dedicated fusion components. The method is designed to capture both modality-specific characteristics and complementary information shared across modalities.
- MulT Tsai et al. (2019) is a transformer-based multimodal fusion framework that models cross-modal dependencies using directional cross-modal attention. For one target modality, the representation of another modality is used as the key and value in the attention operation, allowing the target representation to incorporate information from the source modality.

B.2 Bimodal fusion baselines

For the additional EHR–ECG and EHR–CXR experiments, we compared the proposed model with DAFT, MMTM, MedFuse, and HyperFusion Wolf et al. (2022); Joze et al. (2020); Hayat et al. (2022); Duenias et al. (2025).

- DAFT Wolf et al. (2022) performs dynamic affine feature transformation by predicting sample-specific scaling and shifting parameters from one modality and applying them to the features of another modality. This allows the feature representation of one modality to be conditioned on contextual information from the other modality.
- MMTM Joze et al. (2020) exchanges information across modalities using global feature descriptors. Modality-specific channel activation weights are generated from a joint multimodal representation and are used to recalibrate the features of each modality.
- MedFuse Hayat et al. (2022) was developed to combine clinical time-series data with chest radiographs while accommodating differences in data availability and temporal structure. It independently encodes the input modalities and integrates their representations through a multimodal fusion module for clinical prediction.
- HyperFusion Duenias et al. (2025) uses cross-modal information to dynamically generate or modulate fusion parameters. This mechanism allows the contribution and transformation of one modality to be conditioned on the representations of the other modality.

C Training and model hyperparameters

The proposed framework was implemented in PyTorch and trained on a NVIDIA RTX PRO 6000 Blackwell GPU. It was evaluated on trimodal inputs consisting of EHR, ECG, and CXR. The task was formulated as binary classification of ICU length of stay greater than three days. The dataset was divided into training, validation, and test sets using a stratified 6:2:2 split at the patient level. Three encoder architectures were considered for each modality: FT-Transformer, TabTransformer, and TPC for EHR; WaveNet, 1D-ResNet, and LSTM for ECG; and ResNet-18, DenseNet-121, and VGG-16 for CXR.

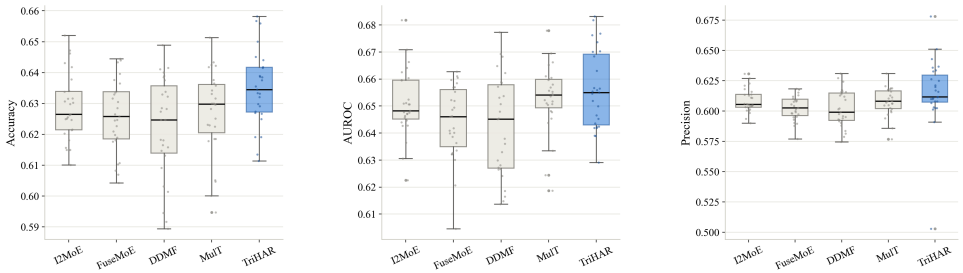
All models were trained using the AdamW optimizer with early stopping based on the validation AUROC of the final prediction output. Training was terminated when the validation AUROC did not improve for 10 consecutive epochs. Gradient clipping and label smoothing were applied throughout training. Each experiment was repeated three times using different random seeds, and the reported performance is presented as the mean and standard deviation across the three runs.

Table 5: Main hyperparameter settings used for training the proposed trimodal framework.

Hyperparameter	**Setting**
Shared representation dimension	64
Standardized sequence length	64
Number of hierarchical levels	3
Gumbel–Softmax temperature	0.7
Number of modality experts	3
Number of interaction experts	3
Number of attention heads	4
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	1×10^{-5}
Batch size	256
Maximum number of epochs	50
Early-stopping patience	10
Label smoothing	0.05
Gradient-clipping norm	5.0
Modality auxiliary-loss weight (λ_{mod})	0.3
Interaction auxiliary-loss weight (λ_{int})	0.3
Gate KL regularization weight	1×10^{-3}

D Additional Experiments

D.1 Trimodal Baseline Comparison: Per-Configuration Distribution



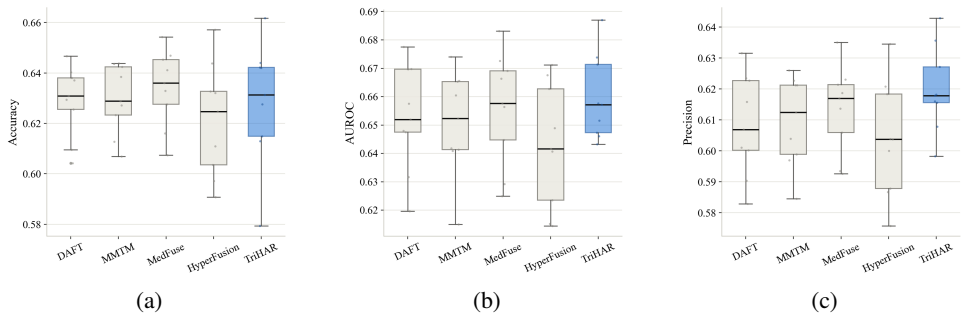
(a) (b) (c)

Figure 5: Distribution of per-configuration (a) Accuracy, (b) AUROC, and (c) Precision across the 27 encoder combinations for TriHAR and the four trimodal baselines. Each point corresponds to one encoder combination, averaged over the three valid seeds. TriHAR is highlighted in blue.

Table 1 in the main text reports the mean performance of TriHAR and the four trimodal baselines (I^2MoE, FuseMoE, DDMF, MulT) for each of the 27 encoder combinations, together with an Average Rank summary. A single averaged rank, however, does not show how consistently a method performs across encoder choices, nor how its performance distribution compares to the alternatives. Figure 5 addresses this by reporting the full distribution of per-configuration means for Accuracy, AUROC, and Precision, with each of the 27 encoder combinations shown as an individual point.

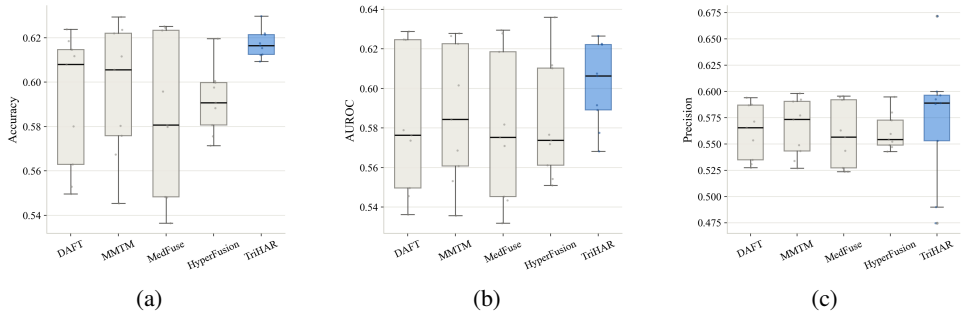
Across all three metrics, TriHAR's distribution is centered at a higher median than every baseline and its interquartile range overlaps only partially with the baselines' upper quartiles, indicating that TriHAR's advantage over the trimodal baselines in Table 1 is not driven by a small number of favorable encoder combinations. MulT is the strongest of the four baselines and its distribution is the closest to TriHAR's, consistent with its ranking in Table 1, while I^2MoE, FuseMoE, and DDMF show wider spreads and lower medians across all three metrics. The overlap that does remain between TriHAR and MulT indicates that for a subset of encoder combinations the two methods are competitive, which is consistent with the encoder-combination-level analysis in Table 1 rather than a contradiction of it.

D.2 Bimodal Baseline Comparison: Per-Configuration Distribution



(a) (b) (c)

Figure 6: Distribution of per-configuration (a) Accuracy, (b) AUROC, and (c) Precision across the 9 EHR–CXR encoder combinations (ECG unavailable) for TriHAR and the four bimodal baselines (DAFT, MMTM, MedFuse, HyperFusion). Each point corresponds to one encoder combination, averaged over the three valid seeds. TriHAR is highlighted in blue.



(a) (b) (c)

Figure 7: Distribution of per-configuration (a) Accuracy, (b) AUROC, and (c) Precision across the 9 EHR–ECG encoder combinations (CXR unavailable) for TriHAR and the four bimodal baselines (DAFT, MMTM, MedFuse, HyperFusion). Each point corresponds to one encoder combination, averaged over the three valid seeds. TriHAR is highlighted in blue.

Table 2 in the main text reports the mean performance of TriHAR and the four bimodal baselines under the EHR+ECG and EHR+CXR settings, together with an Average Rank summary (2.028 for Accuracy, 2.056 for AUROC, and 2.222 for Precision). As with the trimodal comparison, a single averaged rank does not show how consistently TriHAR wins within each bimodal setting, nor whether the two settings behave similarly. Figures 6 and 7 address this by reporting the full per-configuration distribution for each setting separately.

Under EHR+ECG, TriHAR's advantage is large and consistent: its mean Accuracy, AUROC, and Precision (0.617, 0.601, 0.573) each exceed the strongest baseline's mean (MMTM, at 0.596, 0.587, and 0.565, respectively), and TriHAR attains the best per-configuration value in 7, 5, and 4 of the 9 encoder combinations. Under EHR+CXR, the picture is more mixed: TriHAR's mean Accuracy (0.629) falls slightly below MedFuse's (0.634), winning only 2 of the 9 combinations, whereas TriHAR's mean AUROC (0.660) and Precision (0.620) remain the highest of all five methods, winning 5 and 6 of the 9 combinations, respectively.

This asymmetry is consistent with the routing behavior reported earlier. Under the full trimodal model, CXR already receives the largest share of R1+R2's expert weight (Table 4, 58.8%) and is routed almost entirely to the high representation level (Table 3, exceeding 90% for ResNet-18 and approaching 90% for DenseNet-121; see also Figure 9). The high utilization of CXR may leave less room for adaptive routing to improve Accuracy, which may help explain why MedFuse remains competitive when CXR is available. ECG is the weaker of the two auxiliary modalities, so TriHAR's adaptive selection of representation depth and expert weighting provides a larger and more

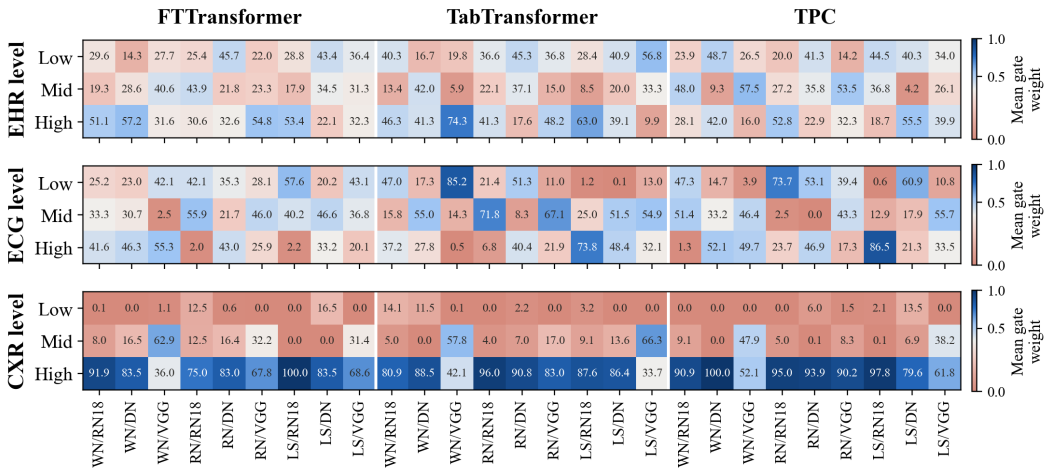


Figure 8: Low/mid/high representation-level gate weights across the 27 encoder combinations, grouped by EHR encoder. Color encodes the mean gate weight relative to the uniform (chance-level) value of $1/3$.

consistent benefit when it is the only modality available besides EHR. Across both settings, TriHAR still attains the highest AUROC and Precision, indicating that its routing benefit is most reliable on ranking-sensitive and positive-class-sensitive metrics even in the harder EHR+CXR setting.

D.3 Representation-Level Routing Analysis

Table 3 in the main text summarizes, for each modality, which representation level (low, mid, or high) TriHAR's first-stage router (R1) selects most often. Figure 8 extends this by reporting the full low/mid/high gate-weight distribution for every one of the 27 encoder combinations, grouped by EHR encoder.

The EHR panel shows architecture-specific depth preferences: FT-Transformer and TabTransformer both assign the largest average share to the high level, while TPC distributes its routing mass more evenly across low, mid, and high. The ECG panel shows a related but distinct pattern: WaveNet and 1D-ResNet both place relatively more weight on low-level representations, while LSTM is comparatively balanced between the mid and high levels. The CXR panel is the most concentrated of the three modalities, with DenseNet-121 and ResNet-18 routed almost entirely to the high level and VGG-16 dividing more of its routing mass between the mid and high levels.

D.4 Expert Utilization Analysis

Table 4 in the main text reports the average EHR/ECG/CXR expert weights assigned by TriHAR's two higher-level routers – R1+R2 (the modality-expert gate) and R1+R3 (the interaction-expert gate). Figure 9 extends this by reporting the full per-configuration expert-utilization heatmap for all 27 encoder combinations, grouped by EHR encoder.

R1+R2 assigns the largest utilization to the direct CXR expert in almost every encoder combination (mean utilization: CXR 58.80%, EHR 22.67%, ECG 18.54%), but the magnitude of this preference varies with the EHR encoder: FT-Transformer combinations tend to allocate a larger share to EHR than the corresponding TabTransformer or TPC combinations, indicating that the modality gate is sensitive to the quality of the underlying direct representation rather than defaulting to a fixed weighting. R1+R3 is comparatively more balanced across its three interaction experts (mean utilization: EHR-enhanced 39.44%, ECG-from-EHR 30.43%, CXR-from-EHR 30.13%), and unlike R1+R2 the leading expert can change across encoder combinations rather than remaining fixed at one modality. This contrast is consistent with the differing roles of the two gates: R1+R2 identifies which modality's direct evidence is strongest, while R1+R3 evaluates how useful each modality's EHR-conditioned interaction representation is, which is a less encoder-dependent quantity.

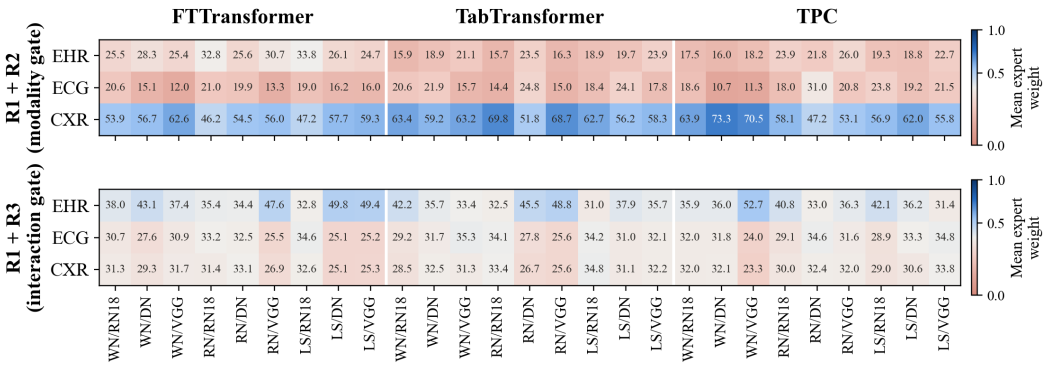
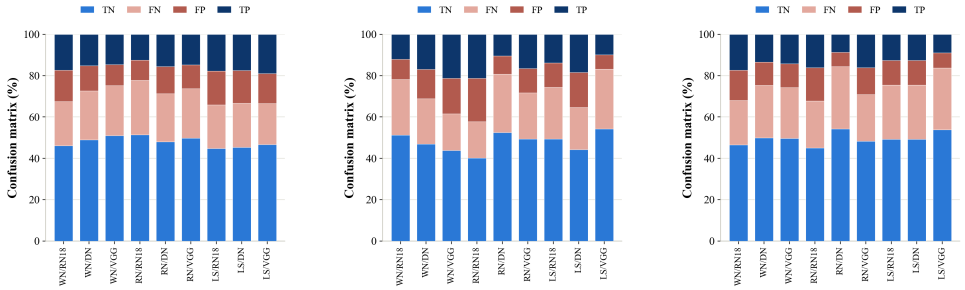


Figure 9: EHR/ECG/CXR expert utilization across the 27 encoder combinations for the R1+R2 (modality-expert) and R1+R3 (interaction-expert) gates. Color encodes the mean expert weight relative to the uniform (chance-level) value of $1/3$.



(a) (b) (c)

Figure 10: Confusion-matrix composition (TN, FN, FP, TP, as a percentage of the $N = 1501$ test set) of the full TriHAR model across the 9 ECG/CXR encoder combinations for each EHR encoder: (a) FT-Transformer, (b) TabTransformer, (c) TPC. TN and TP (correct predictions) form the outer segments of each bar; FN and FP (errors) form the contiguous middle band.

D.5 Confusion Matrix Analysis

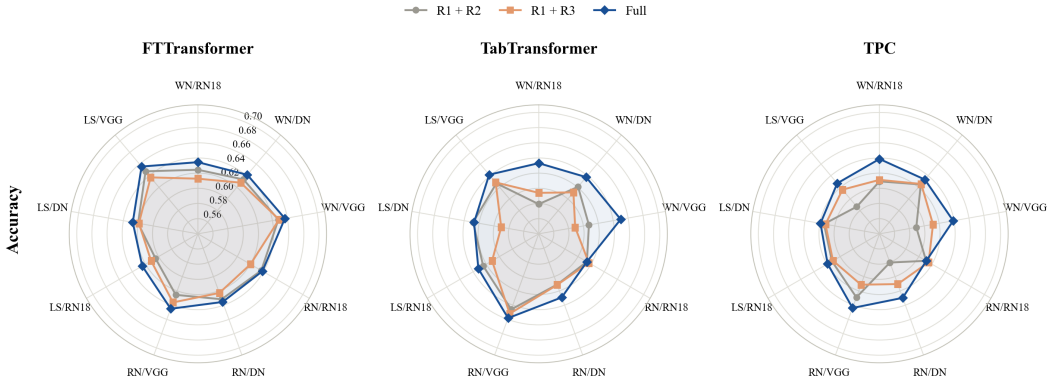
Figure 3 in the main text shows how the confusion-matrix components produced by TriHAR's three routing configurations (R1+R2, R1+R3, and the full model) are distributed across the 27 encoder combinations. Figure 10 provides the complementary, encoder-level view: for the full TriHAR model, it reports the confusion-matrix composition – true negatives (TN), false negatives (FN), false positives (FP), and true positives (TP), each as a percentage of the $N = 1501$ test set – for every individual encoder combination, as a 100%-stacked bar chart, so that a specific configuration's exact operating point can be read off directly rather than inferred from the aggregate distribution in Figure 3. The three panels break this down by EHR encoder.

Within every EHR encoder family, the size of the middle error band (FN + FP) varies noticeably across the 9 ECG/CXR combinations, indicating that encoder choice affects the operating profile of the final classifier even when the aggregate performance values in Table 1 are similar. In several configurations the error band is dominated by false negatives, meaning the model under-identifies long-stay patients and trades recall for a larger true-negative share; in others the split between FN and FP is closer to even. This variation is not concentrated in a single EHR architecture: FT-Transformer, TabTransformer, and TPC each include configurations with a relatively favorable long-stay-detection profile and configurations with stronger short-stay discrimination. No individual ECG–CXR pair dominates all four confusion-matrix entries simultaneously across all three EHR encoders, so Figure 10 should be read together with the aggregate AUROC and Precision results in Table 1 and Figure 3 rather than as evidence for one universally optimal error profile.

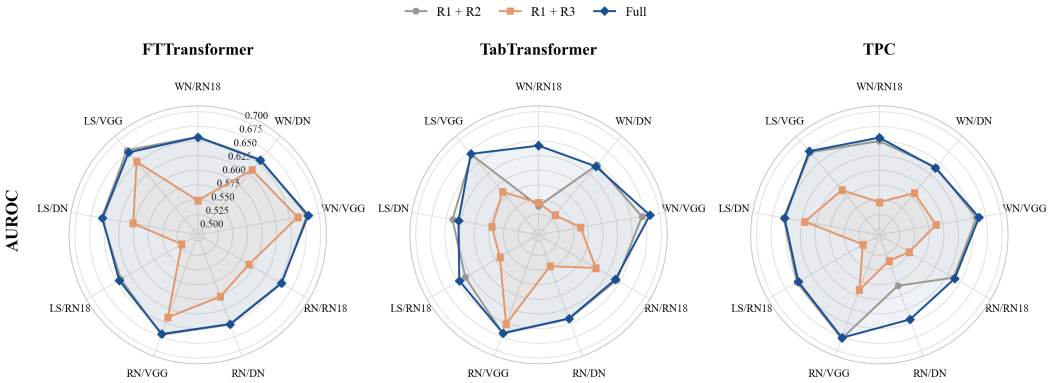
D.6 Routing-Component Performance Analysis

Figure 4 in the main text compares R1+R2, R1+R3, and the full model against the unimodal baselines, pooled across the 27 encoder combinations. Figure 11 reports the complementary per-configuration view: for each metric, a radar chart overlays the three routing configurations – R1+R2, R1+R3, and the full R1+R2+R3 model – as polygons over the 9 ECG/CXR encoder combinations, with one panel per EHR encoder, so that each component's contribution can be read both configuration-by-configuration and as an overall shape.

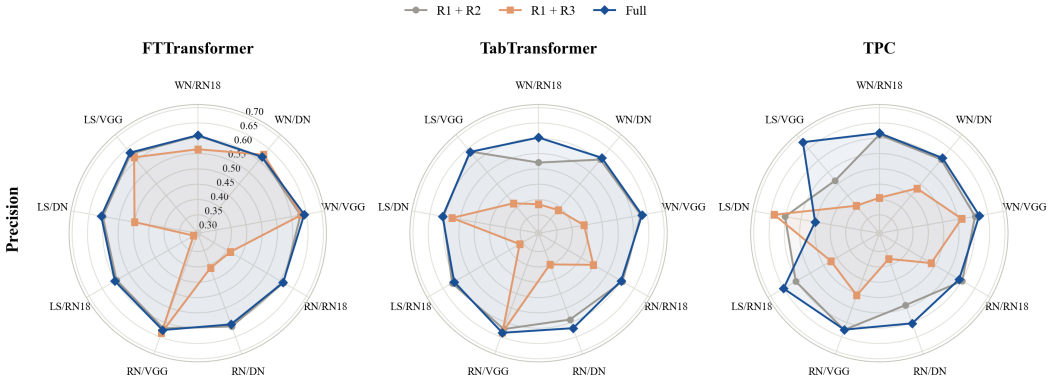
Across all three metrics and all three EHR encoders, the full model's polygon most often encloses, or closely tracks the outer edge of, the R1+R2 and R1+R3 polygons, which is the per-configuration counterpart of its pooled advantage in Figure 4. R1+R2 and R1+R3 trade off which component is closer to the full model depending on metric: for Accuracy and AUROC, R1+R2 tracks the full model more closely across most configurations, while for Precision, R1+R3 is markedly more volatile, producing sharp inward spikes at specific ECG/CXR combinations where the interaction-routing pathway alone is a poor predictor of the positive class, even though the full model remains stable at those same configurations. This indicates that R1+R2 and R1+R3 are not redundant: each captures evidence the other misses, and the full model's consistently larger and smoother polygon reflects the benefit of combining both rather than relying on either routing pathway alone. The variation in polygon shape across FT-Transformer, TabTransformer, and TPC further shows that each routing component's contribution is conditioned on the underlying encoder representation, which motivates rather than undermines the hierarchical adaptive-routing design.



(a)



(b)



(c)

Figure 11: Per-configuration (a) Accuracy, (b) AUROC, and (c) Precision for R1+R2, R1+R3, and the full R1+R2+R3 configuration across the 9 ECG/CXR encoder combinations, with one radar panel per EHR encoder (FT-Transformer, TabTransformer, TPC). The full model is highlighted in blue; identical radial limits are used across the three panels within each metric so that polygon size is directly comparable across EHR encoders.