Multi-Modal Interpretable Graph for Competing Risk Prediction with Electronic Health Records

Munib Mesinovic
munib.mesinovic@jesus.ox.ac.uk
University of Oxford
Oxford, UK

Peter Watkinson
University of Oxford
Oxford, UK

Tingting Zhu
University of Oxford
Oxford, UK

Abstract

We present a novel multi-modal graph learning framework for competing risks survival analysis from electronic health records (EHRs). While recent work has demonstrated the power of deep learning for dynamic risk prediction, most models are constrained to unimodal inputs or static graph structures and cannot model competing clinical endpoints. We introduce a unified, end-to-end model that learns modality-specific spatio-temporal graph representations for time-series, demographics, diagnostic histories, and radiographic text, and fuses them via hierarchical attention into a global patient graph. This design enables dynamic construction of informative substructures both within and across modalities, offering interpretable predictions for multiple competing outcomes. We further propose a composite training objective combining survival likelihood, temporal ranking, and graph regularisation losses to improve risk discrimination, calibration, and structural consistency over time. Our model outperforms state-of-the-art baselines across five real-world EHR datasets, achieving up to 8% gains in cause-specific concordance, while offering fine-grained interpretability across temporal and modality dimensions. These results establish a new foundation for trustworthy and data-efficient risk estimation in clinical settings.

CCS Concepts

• **Applied computing** → **Health informatics**.

Keywords

machine learning for healthcare, survival analysis, graph neural networks, multimodal learning, competing risks, interpretability

ACM Reference Format:
Munib Mesinovic, Peter Watkinson, and Tingting Zhu. 2025. Multi-Modal Interpretable Graph for Competing Risk Prediction with Electronic Health Records. In *Proceedings of Temporal Graph Learning Workshop, SIGKDD International Conference on Knowledge Discovery and Data Mining 2025 (Workshop, KDD 2025).* ACM, New York, NY, USA, 17 pages. https://doi.org/XXXXXXX.XXXXXXX

Workshop, KDD 2025, Toronto, Canada

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The widespread availability of electronic health records (EHRs), combined with recent advances in machine learning, has catalysed the development of clinical risk prediction models [7, 32]. A central challenge is to produce accurate, temporally resolved risk estimates that reflect evolving patient trajectories. Time-to-event modelling addresses this by estimating not only whether an event will occur, but also when, enabling longitudinal risk stratification using sequential data rather than static snapshots [17, 28]. This is especially valuable in acute settings like intensive care, where patient states can rapidly deteriorate [9]. Incorporating multiple clinical modalities—such as vital signs, laboratory measurements, demographics, diagnostic codes, and clinical text—provides a holistic view of patient physiology and has the potential to significantly improve predictive accuracy.

Traditional time-to-event models, such as the Cox proportional hazards model, often assume static covariates and proportional hazards, limiting their flexibility in dynamic and multi-modal contexts [24]. While extensions exist for time-varying covariates and competing risks, they often rely on restrictive assumptions or struggle with complex longitudinal data [2, 40]. Recent deep learning approaches for multi-modal EHR data fuse diverse data streams, but they rarely model competing risks or account for the temporal structure of clinical text and medical histories [10, 31]. To address these gaps, we propose a dynamic multi-modal graph learning framework that builds separate, learnable spatio-temporal graphs for each modality and integrates them with a hierarchical attention mechanism. Our method simultaneously supports interpretability and accurate prediction of multiple competing clinical events.

In this work, we present a new approach to multi-modal learning from EHR data for time-to-event modelling and advance the state-of-the-art by:

- Introducing the first unified framework for dynamic cross-modal graph learning in healthcare, where modality-specific spatio-temporal graphs are constructed and fused through hierarchical attention without relying on predefined structures.
- Proposing a novel hierarchical interpretability mechanism that enables fine-grained attribution across features, time steps, and modalities, under competing events.
- Demonstrating that our approach outperforms state-of-the-art survival models and graph baselines on multiple real-world EHR datasets, achieving calibrated and interpretable competing risk prediction across ICU, emergency, and transplant care.

We evaluated the proposed model in four real-world healthcare datasets from the intensive care and liver transplant settings, and a

survival benchmark dataset, demonstrating superior performance over state-of-the-art methods in deep learning for competing risk prediction while providing robust interpretability.

2 RELATED WORK

2.1 Time-to-event Modelling

Models for time-to-event prediction under competing risks span classical statistical methods and modern deep learning approaches. Traditional techniques such as the Cox proportional hazards model, the Fine–Gray subdistribution model, and Random Survival Forests offer interpretable estimates but rely on strong assumptions like proportional hazards and are ill-suited for high-dimensional, longitudinal EHR data [19, 37]. Discrete-time approaches, including logistic hazard models, provide greater flexibility and are naturally compatible with neural network architectures, though they are often constrained to static inputs and single-risk settings [41]. Deep learning methods such as DeepSurv, DeepHit, Dynamic-DeepHit, and DySurv relax many of these assumptions, enabling direct modelling of cumulative incidence functions and the integration of complex temporal patterns [11, 17, 18, 26]. In ICU contexts, dynamic models like Dynamic-DeepHit and DySurv have demonstrated strong predictive performance by capturing evolving physiological trends and non-linear feature interactions, albeit with reduced interpretability and transparency [29]. Except for Dynamic-DeepHit, most models in EHR provide only independent, cause-specific risk estimates, failing to model competing events jointly. Furthermore, learning from longitudinal and multi-modal EHR data remains challenging due to high dimensionality, irregular sampling, and pervasive missingness, contributing to poor generalizability. To date, no model has explicitly addressed the integration of multiple data modalities and the modelling of competing risks in a unified framework, two critical challenges in machine learning for healthcare. Without the capacity to capture inter-modal and temporal dependencies, existing approaches may fall short in representing patient state accurately for complex prognostic tasks.

2.2 Graph Neural Networks

Graph Neural Networks (GNNs) are well-suited to modeling temporal and structural dependencies in longitudinal health data, particularly in the presence of missingness and modality heterogeneity [3, 35]. Spatio-temporal graphs can capture evolving interactions across features and modalities [36]. Early approaches such as GCNs, GATs, and GraphSAGE rely on pre-defined graphs for downstream tasks [8, 13, 34], constructed via heuristics like Dynamic Time Warping (SimTSC) or convolutional embeddings (MedGNN) [5, 39]. However, such fixed graphs may not reflect task-relevant dynamics. Recent models like TodyNet and DynaGraph remove this constraint by learning graph structures directly from data in an end-to-end manner [23, 25]. While TodyNet struggles with scalability and interpretability, DynaGraph introduces limited temporal interpretability but remains unimodal.

MM-STGNN represents a recent effort to integrate multi-modal data such as demographics, time-series, and imaging for hospital readmission prediction using a shared spatio-temporal graph [33]. However, it fuses modalities via a simple MLP, which underutilises complex temporal and inter-modality dependencies and constructs the graph prior to training using Gaussian kernel similarity on static features, thereby assuming a shared similarity metric across heterogeneous data types. This limits its ability to learn modality-specific dynamics, adapt graph structure during training, or capture fine-grained inter-feature relationships. Moreover, its interpretability and flexibility in modelling complex dependencies are constrained by its reliance on fixed graph topology and global aggregation via GraphSAGE.

3 Data

We evaluate our model on five real-world EHR datasets spanning ICU, emergency department (ED), and transplant settings. For all datasets, laboratory features missing in over 75% of encounters are removed, and remaining missing values are imputed via forward and backward filling. We require at least six time-steps per patient, padding shorter sequences with the most recent observation. Patients are split 8:1:1 into training, validation, and test sets.

MC-MED is a multimodal ED dataset comprising 118,385 visits from 70,545 adults at Stanford Health Care, with time-series vitals, demographics, ICD9/10 histories, labs, and radiography reports. Outcomes include ICU admission (2.3%), hospitalisation (24.9%), and ED observation (13.4%). PBC2 contains monthly longitudinal biomarkers from a primary biliary cirrhosis trial, with death and liver transplantation as competing risks. MIMIC-IV and eICU datasets contain high-resolution ICU data and use in-ICU mortality as the target event; preprocessing follows Mesinovic et al. [26]. The SUPPORT dataset includes static features for 8873 seriously ill inpatients followed for 6-month mortality. Full details are in Supplementary Section A.

4 METHODS

4.1 Notation

Given multi-modal clinical data for P patients, the data for each patient i are represented as a collection of modality-specific feature sets, denoted as $X_i = \left\{X_i^{(\mathrm{T})}, X_i^{(\mathrm{S})}, X_i^{(\mathrm{R})}, X_i^{(\mathrm{C})}\right\}$, where $X_i^{(\mathrm{T})} \in \mathbb{R}^{d_T \times l}$ denotes multivariate time-series features, $X_i^{(\mathrm{S})} \in \mathbb{R}^{d_S}$ denotes static features, $X_i^{(\mathrm{R})} \in \mathbb{R}^{d_R}$ denotes radiographic report embeddings, and $X_i^{(\mathrm{C})} \in \mathbb{R}^{d_C}$ denotes diagnostic code history.

The entire cohort is denoted as $\boldsymbol{X} = \{X_1, X_2, \ldots, X_P\}$. The associated labels are given as $\boldsymbol{Y} = \{Y_1, Y_2, \ldots, Y_P\}$, where each Y_i is a tuple $Y_i = (\epsilon_i, t_i)$. The scalar $t_i \in [0, \infty)$ represents the time observed for an event, defined as $t_i = \min(T_i, C_i)$, where T_i is the true event time and C_i is the censoring time for patient i. The event indicator $\epsilon_i \in \mathcal{E}$, where $\mathcal{E} = \{\varnothing, 1, 2, \ldots, E\}$, denotes the type of observed event, with $\varnothing$ indicating right-censoring (i.e., the absence of an event during the observation window). We assume that censoring is non-informative.

4.2 Dynamic Graph Construction for Time-Series

For longitudinal time-series data, we partition each patient's sequence into s equally-sized time windows. At each time window $t \in \{1, \ldots, s\}$, we construct a dynamic graph where nodes correspond to the f time-series features. The adjacency matrix for

window t, denoted $A_t^{(\mathrm{T})} \in \mathbb{R}^{f\times f}$, is computed using learnable node embeddings:

$$\tilde{A}_t^{(\mathrm{T})} = \Theta_t^\top \cdot \Psi_t$$

where $\Theta_t, \Psi_t \in \mathbb{R}^{d\times f}$ are matrices of learnable embeddings for the source and target nodes. To reduce temporal noise and encourage stability in the evolving graph structure, we apply exponential moving average (EMA) smoothing to the adjacency matrix:

$$A_t^{(\mathrm{T})} = \lambda A_{t-1}^{(\mathrm{T})} + (1-\lambda)\tilde{A}_t^{(\mathrm{T})}, \quad \text{with } A_1^{(\mathrm{T})} = \tilde{A}_1^{(\mathrm{T})}$$

where $\lambda \in [0,1]$ is a smoothing hyperparameter (selected through validation). This encourages gradual transitions between graph topologies over time, improving model convergence and reducing overfitting to local noise.

To promote computational efficiency, each $A_t^{(\mathrm{T})}$ is sparsified by retaining only the top-k highest-weighted edges per node. Specifically, for each node u, we retain edges to the k nodes v with the largest weights $A_{uv}^{(\mathrm{T})}$, excluding self-loops:

$$A_{uv}^{(\mathrm{T})} = 0 \quad \text{if } (u,v) \notin \mathrm{TopK}(A_t^{(\mathrm{T})}, k), \quad u \neq v$$

To explicitly model the evolution of each feature over time, we introduce temporal connections between nodes across consecutive time windows. At each time step t, we instantiate nodes for the current and previous window:

$$\left\{v_{(t,1)}, \ldots, v_{(t,f)}, v_{(t-1,1)}, \ldots, v_{(t-1,f)}\right\}$$

and add directed edges from $v_{(t-1,f)} \rightarrow v_{(t,f)}$ for each feature $f = 1, \ldots, f$, forming a temporally unrolled spatio-temporal graph. To prevent node explosion, embeddings from previous windows are aggregated, and redundant nodes are pruned. The full dynamic graph is represented as a sequence of adjacency matrices:

$$\mathcal{A}^{(\mathrm{T})} = \left\{A_1^{(\mathrm{T})}, A_2^{(\mathrm{T})}, \ldots, A_s^{(\mathrm{T})}\right\} \in \mathbb{R}^{f\times f\times s}$$

capturing intra-window feature dependencies and inter-window temporal transitions.

4.3 Multi-Modal Graph Construction

We use $\mathcal{A}^{(m)}$ to denote the set of modality-specific adjacency matrices, such as dynamic sequences $\mathcal{A}^{(\mathrm{T})} = \{A_1^{(\mathrm{T})}, \ldots, A_s^{(\mathrm{T})}\}$, or single pooled graphs $A^{(m)}$ for static modalities (e.g., static, radiographic, or ICD). Our model simultaneously processes multiple heterogeneous data modalities, constructing an independent graph for each. Specifically, we generate graphs for multivariate time-series measurements, static demographic variables, radiographic report embeddings (X-ray text), and diagnostic code histories (ICD9/10). Each modality is represented by its own adjacency matrix $A^{(m)} \in \mathbb{R}^{d_m\times d_m}$, with optional pooling applied to high-dimensional modalities to reduce the number of nodes and align the input sizes.

Let $C = \{c_1, \ldots, c_{d_C}\}$ be the top $d_C = 500$ most frequent diagnostic codes in the training set. Each code c_i is embedded into a feature vector $e_i \in \mathbb{R}^d$ using a co-occurrence-based embedding model trained on ICD sequences from training data only. We construct a fully connected intra-modality graph $A^{(\mathrm{C})} \in \mathbb{R}^{d_C\times d_C}$ using pairwise cosine similarity:

$$A_{ij}^{(\mathrm{C})} = \frac{e_i^\top e_j}{\|e_i\| \cdot \|e_j\|}, \quad \forall i,j$$

To reduce computational complexity and align the dimensions of the modality across patients, we apply learnable graph pooling to the diagnostic and radiographic graphs. For diagnostic code history $X^{(\mathrm{C})}$, we use a soft-assignment pooling operator $f_\theta^{(\mathrm{C})}$, parameterised as a two-layer neural network, to reduce the entire graph $A^{(\mathrm{C})} \in \mathbb{R}^{d_C\times d_C}$ to a pooled graph of fixed size 50:

$$\tilde{X}^{(\mathrm{C})}, \tilde{A}^{(\mathrm{C})} = f_\theta^{(\mathrm{C})}(X^{(\mathrm{C})}, A^{(\mathrm{C})}),$$

where $\tilde{X}^{(\mathrm{C})} \in \mathbb{R}^{B\times 50\times d}$ and $\tilde{A}^{(\mathrm{C})} \in \mathbb{R}^{B\times 50\times 50}$. Similarly, a pooling module $f_\theta^{(\mathrm{R})}$ is applied to radiographic report graphs $A^{(\mathrm{R})}$, producing fixed-size pooled graphs $\tilde{A}^{(\mathrm{R})} \in \mathbb{R}^{B\times 50\times 50}$. Here, B denotes the batch size and d is the dimensionality of each node's feature vector. Each patient in the batch has a separate pooled graph representation, enabling efficient batched graph computation across modalities.

The pooled graphs $\tilde{A}^{(\mathrm{C})}$ and $\tilde{A}^{(\mathrm{R})}$ are repeated across all s time windows and modulated by modality-specific interpretability matrices $I^{(m)}$, before being incorporated into the final fused adjacency matrix $A^{\mathrm{Fused}} \in \mathbb{R}^{s\times d_{\mathrm{Fused}}\times d_{\mathrm{Fused}}}$.

For radiographic reports, textual feature embeddings are generated using Clinical-Longformer, a transformer model pre-trained on MIMIC-III chest radiograph reports [20, 21]. Let $\mathcal{R}^{(i)} = \{r_1^{(i)}, \ldots, r_T^{(i)}\}$ denote the set of radiographic reports for patient i, ordered by time. Each report is encoded via the frozen Clinical-Longformer model Φ:

$$x_t^{(i)} = \Phi(r_t^{(i)}), \quad x_t^{(i)} \in \mathbb{R}^d$$

We construct a patient-specific temporal graph $A^{(i,\mathrm{R})} \in \mathbb{R}^{T\times T}$ using a Gaussian similarity kernel:

$$A_{tt'}^{(i,\mathrm{R})} = \exp\left(-\frac{\|x_t^{(i)} - x_{t'}^{(i)}\|_2^2}{\tau}\right), \quad \forall t,t' \in \{1,\ldots,T\}$$

where τ is a temperature hyperparameter. The corresponding node feature matrix is $X^{(i,\mathrm{R})} = [x_1^{(i)}, \ldots, x_T^{(i)}]^\top$.

Each intra-modality graph $A^{(m)}$ (including TS, Static, ICD, Xray) is paired with a corresponding interpretability matrix $I^{(m)} \in \mathbb{R}^{d_m\times d_m}$ as described below. These matrices are trained jointly with the model and updated based on gradient attribution to quantify feature-level and structural importance for survival prediction. Together, these modality-specific graphs contribute to the final fused multi-modal graph A^{Fused} and its interpretability matrix I^{Fused} used in downstream graph learning.

4.4 Hierarchical Attention for Multimodality

To enable interactions between modalities, we introduce learnable cross-modality attention matrices, $W^{(m\to n)} \in \mathbb{R}^{d_m\times d_n}$, for each pair of ordered modality (m,n), where $W^{(m\to n)}$ captures the directed influence from the features of the modality m to those of the modality n. These matrices are randomly initialised and optimised end-to-end to support the survival prediction objective. Each intra-modality graph $A^{(m)} \in \mathbb{R}^{d_m\times d_m}$ and cross-modality matrix

$W^{(m\to n)} \in \mathbb{R}^{d_m \times d_n}$ is paired with a corresponding interpretability matrix: $I^{(m)} \in \mathbb{R}^{d_m \times d_m}$ and $I^{(m\to n)} \in \mathbb{R}^{d_m \times d_n}$, respectively. These matrices quantify the saliency of each node or interaction and are updated via gradient-based attribution with respect to the model loss (see Supplementary Section 3 for computation and Section 4 for convergence). The fused multi-modal graph is assembled by concatenating intra- and inter-modality blocks:

$$A^{\text{Fused}} = \begin{bmatrix} A^{(\text{T})} & W^{(\text{T}\to\text{S})} & W^{(\text{T}\to\text{R})} & W^{(\text{T}\to\text{C})} \\ W^{(\text{S}\to\text{T})} & A^{(\text{S})} & W^{(\text{S}\to\text{R})} & W^{(\text{S}\to\text{C})} \\ W^{(\text{R}\to\text{T})} & W^{(\text{R}\to\text{S})} & A^{(\text{R})} & W^{(\text{R}\to\text{C})} \\ W^{(\text{C}\to\text{T})} & W^{(\text{C}\to\text{S})} & W^{(\text{C}\to\text{R})} & \tilde{A}^{(\text{C})} \end{bmatrix} \in \mathbb{R}^{d_{\text{Fused}} \times d_{\text{Fused}}}$$

where diagonal blocks denote intra-modality adjacency matrices and off-diagonal blocks encode directed inter-modality interactions. The interpretability matrices are aggregated into a fused matrix $I^{\text{Fused}} \in \mathbb{R}^{d_{\text{Fused}} \times d_{\text{Fused}}}$ with a matching block structure:

$$I^{\text{Fused}} = \begin{bmatrix} I^{(\text{T})} & I^{(\text{T}\to\text{S})} & \cdots \\ \vdots & \ddots & \\ & & I^{(\text{C})} \end{bmatrix}$$

which provides global attention weights across all node pairs. The final fused graph G^{Final} passed to downstream GNN layers is computed by element-wise modulation:

$$G^{\text{Final}} := A^{\text{Fused}} \circ \text{softmax}(I^{\text{Fused}})$$

where $\circ$ denotes the Hadamard product and softmax is applied row-wise to normalise attention weights. This hierarchical attention mechanism enables both fine-grained feature-level and modality-level interpretability. Further implementation details on Graph Isomorphism Network (GIN) learning are provided in Supplementary Section 5.

4.5 Model Training

Our hierarchical graph fusion framework is explicitly designed to support inference under modality-specific missingness, a common challenge in real-world clinical settings. During training, the model learns separate modality-specific graphs (e.g., time-series, ICD codes, radiography, static features) and fuses them via hierarchical attention into a global multi-modal graph representation. At inference time, if a particular modality is unavailable (e.g., no radiography report or no ICD history), the corresponding graph $A^{(m)}$ and its inter-modality attention matrices $W^{(m\to n)}$ and $W^{(n\to m)}$ are excluded from the fusion process. This results in a reduced fused graph A^{Fused} composed only of the available modalities. The interpretability matrix I^{Fused} is pruned similarly to reflect the updated block structure. Since the GNN operates on this fused graph regardless of its size, the model can still propagate information and compute patient-specific risk estimates using only the available data sources, without requiring retraining or imputation.

Due to the complexity of the model, we use temporal pooling on the learned graphs from the GIN layers as described in Supplementary Section 6, and we add a regularisation term to help prevent overfitting and support learning stability:

$$\mathcal{L}_{\text{reg}} = \lambda \sum_{(k,l)\in E} \|\mathbf{h}_k - \mathbf{h}_l\|^2 \tag{1}$$

where λ is a hyperparameter controlling the strength of the regularization, (k, l) represents an edge connecting nodes k and l, and $\mathbf{h}_k, \mathbf{h}_l$ are the feature representations of nodes k and l, respectively. Please note that E here represents the set of edges in the dynamically constructed graph, where each edge connects a pair of nodes (features).

For our right-censored setup, we need to estimate the joint distribution of the first event time and multiple competing events. We do so by directly estimating the negative log-likelihood of this distribution. For those patients who have suffered the specific event, we capture both the outcome and the time at which it occurs. For censored patients, we capture the censoring time conditioned on the measurements recorded prior to the censoring. If we assume $\widehat{a}_t = \hat{P}(T = \delta, \epsilon = \epsilon \mid X)$ represents the estimated probability of experiencing an event ϵ at time δ, this loss is defined as:

$$\begin{aligned} \mathcal{L}_{NLL} = -\sum_{i=1}^{N} \Bigg[& \mathbb{1}\left(\epsilon^i \neq \varnothing\right) \cdot \log\left(\frac{\widehat{a}^i_{\epsilon^i,\delta^i}}{1 - \sum_{\epsilon\neq\varnothing} \sum_{n \le t^i_{ji}} \widehat{a}^i_{\epsilon,n}} \right) \\ & + \mathbb{1}\left(\epsilon^i = \varnothing\right) \cdot \log\left(1 - \sum_{\epsilon\neq\varnothing} \hat{F}_\epsilon\left(\delta^i \mid X^i\right)\right) \Bigg] \end{aligned} \tag{2}$$

The second component of the loss is a ranking loss designed to refine the model's ability to estimate cause-specific cumulative incidence functions (CIFs). Inspired by Lee et al. [17], this loss encourages the model to assign higher risk scores at earlier event times for patients who experience the event sooner than others who survive longer. Because patients' longitudinal measurements may begin at different stages of their clinical course, however, direct comparisons based on absolute event times may not be meaningful. To address this, we compute relative times with respect to each patient's most recent measurement, defining $s_i = \delta_i - t_i^{(J_i)}$, where δ_i is the event time and $t_i^{(J_i)}$ denotes the timestamp of the last available observation for subject i. To preserve the causal structure of time-to-event modelling, we ensure that the model only uses information available up to the prediction time t when estimating the CIF. Specifically, input node features (e.g., time-series, ICD codes, radiography embeddings) are truncated to timestamps $t' \le t$, and the graph structure is dynamically constructed based on the clinical state at each t. This design prevents any leakage of future information. Although the model is trained to predict cumulative event probabilities over a future horizon $[t, T]$, all feature encoding and message passing are strictly limited to historical data. This ensures temporally valid training and enables real-time inference in clinical settings.

We then define a pair (i, j) as an acceptable pair for event type ϵ if subject i experiences event ϵ at time s_i, while subject j does not experience any event until s_i, i.e., $s_j > s_i$. For these pairs, the model's predicted CIFs should satisfy:

$$\hat{F}_\epsilon\left(s_i + t_i^{(J_i)} \mid X_i\right) > \hat{F}_\epsilon\left(s_i + t_j^{(J_j)} \mid X_j\right)$$

ensuring that the estimated risk for subject i, who experienced the event earlier, is higher than that for subject j, who survived beyond that point. The ranking loss is computed over such acceptable pairs, allowing the model to learn consistent risk ordering across patients with heterogeneous measurement histories. For notational

simplicity, let $R_\epsilon^{(i)} := \hat{F}_\epsilon\left(s_i + t_i^{(J_i)} \mid X_i\right)$ denote the predicted cumulative incidence for event type ϵ at the effective event time for subject i. The ranking loss is then:

$$\mathcal{L}_{rank} = \sum_{\epsilon=1}^{E} \mu \sum_{i \neq j} M_{ij}^{\epsilon} \cdot \eta\left(R_\epsilon^{(i)}, R_\epsilon^{(j)}\right) \tag{3}$$

where $M_{ij}^{\epsilon} = 1\left(\epsilon_i = \epsilon, s_i < s_j\right)$ indicates whether the pair (i, j) is acceptable for event type ϵ, and $\eta(\cdot)$ is a smooth loss function comparing the predicted CIFs. We adopt the formulation $\eta(a, b) = \exp\left(-\frac{a-b}{\sigma}\right)$, which penalises incorrect risk rankings in a margin-sensitive manner. For simplicity, we assume uniform weighting across event types with $\mu_\epsilon = \mu$ for all ϵ. This ranking term is integrated into the total loss to improve temporal discrimination across competing risks, particularly in the presence of censored and irregular longitudinal data.

The final loss is the sum of the regularisation loss for complexity and overfitting adjustment, the joint negative log-likelihood of the competing risks, and a ranking loss to fine-tune the model to each cause accordingly. α, β, and γ are considered hyperparameters:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{reg}} + \alpha \mathcal{L}_{\text{NLL}} + \beta \mathcal{L}_{\text{rank}} \tag{4}$$

After creating the final graph embedding through graph expansion and overlaying hierarchical attention, G_s is passed through a multi-layer temporal graph isomorphism network (GIN) for graph learning. The GIN output is a graph pooled temporally to reduce the number of nodes with convolutional clustering and decrease the computational costs of the training. The reduced graph is then flattened and passed through E parallel multilayer perception subnetworks and a joint softmax layer to produce measures of the probability of the estimated joint distribution of event times and competing events. For a patient X, since the event times are discretised, the estimated CIF for a specific cause at time δ is:

$$\hat{F}_\epsilon\left(\delta \mid X\right) = \frac{\sum_{t_J < \delta \leq \Delta} \widehat{a}_{\epsilon,\delta}}{1 - \sum_{\epsilon \neq \varnothing} \sum_{n \leq t_J} \widehat{a}_{\epsilon,n}} \tag{5}$$

5 RESULTS

5.1 Model Comparison

We evaluate our proposed model against two groups of baselines under both single-risk and competing-risk survival analysis scenarios. The first group comprises state-of-the-art statistical and deep learning models for survival analysis, including static models (Cox PH, DeepSurv, DeepHit) and dynamic models that can process longitudinal time-series data (Dynamic-DeepHit, DySurv). The second group consists of models of graph neural networks adapted for survival tasks, including GCN, GAT, GraphSAGE, TodyNet, DynaGraph, and MM-STGNN. We include the discrete logistic hazard loss from DeepHit and DySurv as the objective for these models to allow them to approximate time-to-events. For MM-STGNN, we replace the original imaging modality with textual embeddings from radiographic reports to match our input setup and use the competing risk loss function for survival adaptation, similarly for MedGNN. For TodyNet and DynaGraph, we use the time-series and/or static modalities from the relevant datasets. Details on experimental settings and implementation choices are provided in Supplementary Section 1.

All experiments were conducted using PyTorch 3.8 and an NVIDIA A100 Tensor Core GPU. Evaluation is based on three complementary metrics: time-dependent concordance (including cause-specific concordance for competing risks), integrated Brier score (IBS), and integrated binomial log-likelihood (IBLL), as detailed in Supplementary Section 2. At inference time, the model receives patient data up to the desired prediction horizon and outputs the cumulative incidence estimate for that horizon.

A complete set of comparisons visualised as spider plots across models and metrics is available in Supplementary Section 7 for single-risk settings (hospital admission for MC-MED and death for PBC2). Table 1 shows the results in the same settings as the results averaged over five seeds. Our model consistently outperforms all baselines across different datasets and care environments. The largest improvements are observed in concordance metrics, reflecting our model's strong temporal discrimination capabilities. The combination of ranking loss and regularisation enables our model to predict both short- and long-term survival outcomes robustly across diverse clinical settings. Moreover, the superior IBS and IBLL scores compared to Dynamic-DeepHit and related methods indicate better generalisation and calibrated event probability estimates, avoiding the inflated concordance behaviour seen in prior ranking-based models.

We evaluated our proposed model against competing risk prediction, including Dynamic-DeepHit, MedGNN, and MM-STGNN. As shown in Table 2, our model outperforms all baselines across both the PBC2 and MC-MED datasets in terms of cause-specific time-dependent concordance. In PBC2, our model improves over MedGNN and MM-STGNN by 2.5% and 2.0%, respectively, in predicting death, with similar gains for liver transplant prediction. In MC-MED, the advantage is most notable in predicting hospital admissions, where our model achieves concordance 0.880 ± 0.010, outperforming MM-STGNN (0.798 ± 0.012) and MedGNN (0.791 ± 0.012). These results highlight the advantage of our dynamic and hierarchical graph formulation, which jointly learns intra- and inter-modality structure with attention-based interpretability, in contrast to fixed similarity graphs or late fusion approaches used in prior work.

5.2 Model Interpretability

To better understand the temporal dynamics of model decision-making, we visualise the interpretability weights associated with the top 10 time-series features over six time steps for each of the three competing outcomes in the MC-MED dataset. These weights, derived from the modality-specific interpretability matrix $I^{(\mathrm{T})}$ (see Section 4.4), are normalised and plotted as heatmaps in Figure 1. Each heatmap reveals how feature importance evolves throughout the stay in the emergency department (ED) of a patient, with each time step corresponding to a discrete and regular interval in the model prediction window. Higher magnitudes indicate a stronger contribution to model loss during training, and hence to the predicted cumulative incidence function (CIF) for each outcome.

Table 1: Single-risk evaluation results across five datasets using three time-to-event metrics. Time-dependent concordance index (↑): higher is better; integrated Brier score (IBS, ↓) and integrated binomial log-likelihood (IBLL, ↓): lower is better. Values are reported as mean (standard deviation) over 5 seeds. Best values per dataset are in bold.

Concordance Index (↑)

Model	MIMIC-IV	eICU	PBC2	MC-MED	SUPPORT
Cox PH	0.711 (0.013)	0.642 (0.016)	0.676 (0.012)	0.589 (0.018)	0.664 (0.014)
DeepSurv	0.752 (0.011)	0.684 (0.014)	0.702 (0.015)	0.654 (0.017)	0.698 (0.013)
DeepHit	0.778 (0.010)	0.723 (0.012)	0.706 (0.011)	0.739 (0.014)	0.719 (0.012)
D-DeepHit	0.807 (0.009)	0.758 (0.010)	0.716 (0.011)	0.786 (0.010)	0.749 (0.011)
DySurv	0.832 (0.008)	0.782 (0.009)	0.736 (0.009)	0.809 (0.008)	0.779 (0.009)
GCN	0.649 (0.014)	0.598 (0.015)	0.631 (0.012)	0.579 (0.015)	0.622 (0.014)
GAT	0.674 (0.012)	0.621 (0.013)	0.659 (0.011)	0.602 (0.014)	0.647 (0.012)
GraphSAGE	0.697 (0.013)	0.643 (0.012)	0.681 (0.010)	0.634 (0.014)	0.669 (0.013)
TodyNet	0.738 (0.011)	0.685 (0.010)	0.706 (0.011)	0.719 (0.010)	0.708 (0.011)
DynaGraph	0.803 (0.008)	0.744 (0.009)	0.726 (0.010)	**0.832 (0.008)**	0.761 (0.010)
MedGNN	0.759 (0.009)	0.708 (0.011)	0.698 (0.010)	0.776 (0.009)	0.732 (0.010)
MM-STGNN	0.767 (0.010)	0.725 (0.010)	0.693 (0.012)	0.762 (0.011)	0.731 (0.011)
Ours	**0.861 (0.007)**	**0.809 (0.008)**	**0.768 (0.008)**	**0.832 (0.007)**	**0.797 (0.008)**

Integrated Brier Score (IBS) (↓)

Model	MIMIC-IV	eICU	PBC2	MC-MED	SUPPORT
Cox PH	0.251 (0.012)	0.304 (0.015)	0.281 (0.014)	0.332 (0.016)	0.270 (0.012)
DeepSurv	0.222 (0.011)	0.261 (0.013)	0.246 (0.013)	0.299 (0.015)	0.238 (0.012)
DeepHit	0.209 (0.010)	0.248 (0.012)	0.257 (0.012)	0.278 (0.013)	0.247 (0.011)
D-DeepHit	0.186 (0.009)	0.219 (0.011)	0.237 (0.011)	0.199 (0.012)	0.227 (0.010)
DySurv	0.171 (0.008)	0.209 (0.010)	0.218 (0.009)	0.177 (0.010)	0.210 (0.009)
GCN	0.321 (0.014)	0.357 (0.016)	0.336 (0.014)	0.374 (0.015)	0.351 (0.014)
GAT	0.301 (0.012)	0.341 (0.014)	0.318 (0.013)	0.351 (0.014)	0.327 (0.013)
GraphSAGE	0.282 (0.013)	0.324 (0.012)	0.309 (0.012)	0.336 (0.014)	0.298 (0.012)
TodyNet	0.239 (0.011)	0.279 (0.010)	0.274 (0.011)	0.257 (0.012)	0.285 (0.011)
DynaGraph	0.198 (0.009)	0.234 (0.010)	0.246 (0.011)	0.216 (0.011)	0.239 (0.010)
MedGNN	0.186 (0.009)	0.215 (0.010)	0.231 (0.010)	0.209 (0.010)	0.227 (0.009)
MM-STGNN	0.202 (0.010)	0.242 (0.011)	0.259 (0.011)	0.212 (0.011)	0.252 (0.010)
Ours	**0.139 (0.006)**	**0.183 (0.007)**	**0.189 (0.007)**	**0.128 (0.007)**	**0.171 (0.006)**

Integrated Binomial Log-Likelihood (IBLL) (↓)

Model	MIMIC-IV	eICU	PBC2	MC-MED	SUPPORT
Cox PH	-0.223 (0.011)	-0.257 (0.012)	-0.243 (0.011)	-0.284 (0.012)	-0.249 (0.011)
DeepSurv	-0.258 (0.010)	-0.299 (0.011)	-0.281 (0.011)	-0.317 (0.011)	-0.288 (0.010)
DeepHit	-0.278 (0.009)	-0.322 (0.010)	-0.304 (0.010)	-0.337 (0.011)	-0.308 (0.010)
D-DeepHit	-0.318 (0.008)	-0.369 (0.009)	-0.351 (0.009)	-0.381 (0.010)	-0.356 (0.009)
DySurv	-0.339 (0.007)	-0.391 (0.009)	-0.367 (0.008)	-0.413 (0.009)	-0.379 (0.008)
GCN	-0.184 (0.012)	-0.212 (0.013)	-0.194 (0.012)	-0.228 (0.013)	-0.202 (0.012)
GAT	-0.201 (0.011)	-0.238 (0.012)	-0.219 (0.011)	-0.251 (0.012)	-0.226 (0.011)
GraphSAGE	-0.215 (0.011)	-0.248 (0.011)	-0.234 (0.010)	-0.271 (0.011)	-0.238 (0.010)
TodyNet	-0.292 (0.010)	-0.335 (0.010)	-0.311 (0.010)	-0.353 (0.010)	-0.319 (0.010)
DynaGraph	-0.298 (0.009)	-0.342 (0.010)	-0.318 (0.010)	-0.362 (0.010)	-0.329 (0.010)
MedGNN	-0.308 (0.009)	-0.362 (0.010)	-0.337 (0.009)	-0.373 (0.009)	-0.347 (0.009)
MM-STGNN	-0.301 (0.010)	-0.354 (0.010)	-0.328 (0.010)	-0.361 (0.010)	-0.338 (0.010)
Ours	**-0.398 (0.006)**	**-0.442 (0.007)**	**-0.417 (0.007)**	**-0.459 (0.007)**	**-0.426 (0.007)**

The interpretability weights exhibit both temporal variability and sparsity, demonstrating the model's ability to identify and adapt to transient but informative physiological signals. For example, respiratory rate (RR) and ambulatory vitals show a high level of attention early for ED observation, while renal markers such as creatinine and BUN become more influential for hospital admission over time. These trends align with clinical intuition: early instability informs immediate triage decisions, whereas lab-based markers accumulate relevance for longer-term outcomes.

5.3 Model Evaluation

5.3.1 Ablation Studies. To assess the contribution of individual components within our proposed model framework, we conducted a comprehensive ablation study across both competing risk and single-risk survival settings. To quantify the marginal utility of each data modality, we conducted a modality addition experiment

Table 2: Cause-specific time-dependent concordance index ($C^{\text{td}}_{\text{ind},\epsilon}$) for competing risk prediction on the PBC2 and MC-MED datasets. On MC-MED, our model uses multi-modal input comprising time-series, static demographics, ICD code history, and radiographic text embeddings. MM-STGNN used time-series, demographics, and radiographic text instead of images. In contrast, Dynamic-DeepHit and MedGNN use only static and time-series input. Values are averaged over five seeds; mean (std) reported. Higher is better.

Dataset	Model	Death	Liver Transplant	–
PBC2	MedGNN	0.758 (0.012)	0.740 (0.015)	–
	Dynamic-DeepHit	0.765 (0.011)	0.743 (0.014)	–
	MM-STGNN	0.770 (0.010)	0.749 (0.013)	–
	Ours	**0.790 (0.009)**	**0.766 (0.010)**	–
	Model	**ICU Admission**	**Hospital Admission**	**ED Observation**
MC-MED	MedGNN	0.812 (0.011)	0.785 (0.014)	0.773 (0.013)
	Dynamic-DeepHit	0.816 (0.009)	0.791 (0.012)	0.778 (0.013)
	MM-STGNN	0.821 (0.010)	0.798 (0.012)	0.781 (0.014)
	Ours	**0.827 (0.007)**	**0.880 (0.010)**	**0.797 (0.012)**

(Table 4) on the MC-MED dataset, incrementally adding time-series, ICD codes, and radiography embeddings to a static only baseline. The results show that time-series features provide the most substantial improvement in cause-specific concordance, while the ICD and radiographic data provide complementary gains, supporting the model's ability to leverage heterogeneous modalities through hierarchical fusion. In the MC-MED dataset, which involves three competing outcomes (ICU admission, hospital admission, and observation in the emergency department), the entire model consistently achieves the highest cause-specific concordance across all outcomes (Table 3), demonstrating the effectiveness of integrating architectural, loss, and input design choices. Removing the ranking loss or the cross-modality attention, each leads to measurable drops in performance, highlighting their complementary roles. The absence of cross-modal attention reduces the model's ability to integrate heterogeneous signals, especially in the hospital admission outcome, which depends more heavily on fused modalities such as radiographic findings and diagnostic history.

Notably, removing the exponential moving average (EMA) smoothing from the dynamic graph construction results in performance degradation across all outcomes, with the most pronounced effect on hospital admission (0.814 vs 0.880). This confirms that temporal stability in the evolving graph structure is crucial for capturing disease progression patterns, particularly for outcomes that develop over longer time horizons. The EMA mechanism effectively reduces noise in the temporal graph transitions, allowing the model to learn more robust representations of patient state evolution. Furthermore, replacing the GIN module with a standard GCN results in consistent declines across outcomes, underscoring the benefit of structure-aware message passing in our dynamic graph encoder.

5.3.2 Out-Of-Distribution and Calibration Experiments. Real-world clinical settings often involve incomplete data due to modality-specific missingness. To assess model robustness in such scenarios, we perform an out-of-distribution (OOD) evaluation in which one modality is removed at inference time without retraining the model. This simulates deployment-time conditions where patients may lack certain inputs, such as missing ICD codes for new admissions,

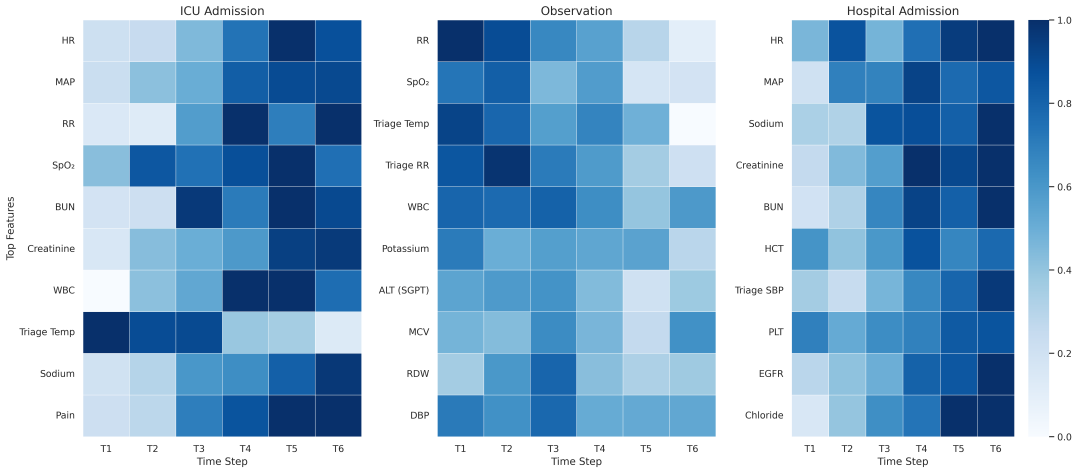


Figure 1: Temporal interpretability heatmaps of the top 10 most predictive time-series features for each of the three competing outcomes in the MC-MED dataset (ICU admission, ED observation, hospital admission). We visualise the average interpretability weights across 5 runs for each feature over 6 time steps. We normalise each matrix row-wise using min-max scaling to ensure comparability across features and outcomes. The attention scores are extracted from the modality-specific interpretability matrix I^{T} for each outcome, highlighting both temporally local and globally consistent predictive contributions.

Table 3: Ablation study on the MC-MED dataset. We report the cause-specific time-dependent concordance index ($C^{\mathrm{td}}_{\mathbf{ind},\epsilon}$) for three competing outcomes (ICU admission, hospital admission, in-hospital observation), averaged over five seeds. Higher is better. Mean (std) shown.

Model Variant	ICU Admission	Hospital Admission	Observation
w/o Ranking Loss	0.802 (0.012)	0.783 (0.015)	0.776 (0.017)
w/o ICD	0.822 (0.009)	0.828 (0.011)	0.785 (0.012)
w/o Radio	0.820 (0.009)	0.820 (0.011)	0.782 (0.012)
w/o EMA	0.811 (0.022)	0.814 (0.027)	0.773 (0.028)
GCN backbone	0.811 (0.008)	0.813 (0.011)	0.782 (0.013)
w/o Hierarchical Attn	0.805 (0.010)	0.812 (0.012)	0.778 (0.013)
Ours (Full Model)	**0.827 (0.007)**	**0.880 (0.010)**	**0.797 (0.012)**

Table 4: Expanded modality ablation study on the MC-MED dataset. We report the cause-specific time-dependent concordance index ($C^{\mathrm{td}}_{\mathbf{ind},\epsilon}$) for three competing outcomes: ICU admission, hospital admission, and in-hospital observation. Results are averaged over five seeds. Mean (std) shown.

Modalities Included	ICU Admission	Hospital Admission	Observation
Static only	0.792 (0.011)	0.769 (0.013)	0.766 (0.014)
Static + TS	0.815 (0.010)	0.790 (0.012)	0.776 (0.013)
Static + ICD	0.798 (0.010)	0.790 (0.012)	0.772 (0.013)
Static + Radio	0.796 (0.011)	0.785 (0.013)	0.770 (0.014)
Static + TS + ICD	0.820 (0.009)	0.820 (0.011)	0.782 (0.012)
Static + TS + Radio	0.822 (0.009)	0.828 (0.011)	0.785 (0.012)
Static + ICD + Radio	0.809 (0.010)	0.829 (0.012)	0.781 (0.013)
All Modalities	**0.827 (0.007)**	**0.880 (0.010)**	**0.797 (0.012)**

absent radiographic reports for milder cases, or unavailable vitals. Table 5 shows the cause-specific concordance index on the MC-MED dataset when individual modalities are masked during testing. Although performance drops relative to the full input model, degradation is gradual and modality-dependent. The largest declines occur when time-series data are removed, highlighting their centrality to dynamic risk estimation. Our model still retains a useful discriminative ability from the remaining modalities, with the prediction of hospital admission maintaining a concordance of 0.808 even without input from time-series.

In addition to ranking-based and discrimination metrics, we evaluate the probabilistic calibration of our model compared to Dynamic-DeepHit across two ICU datasets and three prediction horizons. Figure 3 presents calibration plots at 48h, 96h, and 144h for both the MIMIC-IV (top row) and eICU (bottom row) datasets, alongside bootstrap-based confidence intervals and the reported expected calibration error (ECE) for each model. In both cohorts, our model demonstrates closer alignment with the ideal calibration line and consistently lower ECE values. At the 48h prediction horizon, our model achieves an ECE of 0.024 on MIMIC-IV and 0.030 on eICU, compared to 0.068 and 0.082 for Dynamic-DeepHit, respectively. This performance trend persists across longer horizons as well, indicating that our model not only produces better discriminative rankings but also more calibrated survival probabilities. These results highlight the robustness of the model across care settings and its suitability for clinical risk estimation tasks that require reliable uncertainty quantification.

6 DISCUSSION

Most prior models for survival analysis, including Dynamic-DeepHit, do not consider multi-modality when issuing dynamic and high-risk time-to-event predictions. Prior dynamic graph learning assumes that a static graph representation can sufficiently capture the temporal information required for downstream tasks. Models such as GCNs and GATs often adopt this approach, but they fail to account for potential associations that arise when time-series data are represented by multiple graphs over time [4, 13]. Dynamic graph models such as MedGNN or multi-modal graph models such as MM-STGNN treat modalities as independent input streams or rely on early fusion techniques that fail to capture the full structure

Table 5: Modality-specific OOD robustness on the MC-MED dataset. Each row reports the cause-specific concordance index ($C^{\text{td}}_{\text{ind},\epsilon}$) when one modality is removed at inference time. The model is trained on full inputs and evaluated on partially missing data, simulating real-world deployment with incomplete records. While performance degrades smoothly, it shows robustness to clinical missingness patterns.

Missing Modality at Inference	ICU Admission	Hospital Admission	Observation
None (Full Model)	**0.827 ± 0.007**	**0.880 ± 0.010**	**0.797 ± 0.012**
Radiography/Text	0.819 ± 0.008	0.846 ± 0.011	0.789 ± 0.012
ICD Codes	0.818 ± 0.008	0.841 ± 0.012	0.788 ± 0.013
Time-Series	0.791 ± 0.010	0.808 ± 0.014	0.772 ± 0.014
Static Demographics	0.808 ± 0.009	0.856 ± 0.012	0.785 ± 0.013

of temporal and cross-modal dependencies [5, 33]. Our proposed model addresses these limitations by integrating dynamic graph representations, modality-specific graph learning, and hierarchical attention to deliver interpretable, patient-specific risk predictions under competing risks.

Through extensive benchmarking, we find that our model consistently outperforms existing models on both single-risk and competing-risk prediction tasks across diverse datasets. These improvements are most evident in concordance metrics, where our model demonstrates a superior ability to temporally rank patient outcomes. The integration of a ranking loss with stabilised graph weights proves especially effective, helping the model maintain consistency in its temporal predictions while stabilising graph construction over time. In addition, calibration results across multiple prediction horizons show that the model produces well-calibrated risk estimates, a critical requirement for clinical deployment.

A key strength of our model lies in its ability to learn and leverage the modality-specific structure. Unlike previous models that apply shared similarity metrics across heterogeneous features, it constructs and trains distinct graph representations for time-series, static demographics, diagnostic history (ICD codes), and radiographic text. The hierarchical attention mechanism allows the model to learn cross-modal influence while preserving intra-modality interpretability. Visualisations of both temporal feature weights and modality-level attention reveal clinically significant patterns like the increasing importance of renal markers and blood pressure over time for hospital admissions [14, 38], and early signal from triage vitals, especially initial temperature readings [12], and medical history for ICU escalation. Clinical research on this latter association is currently limited. These insights help clinicians understand not just what the model predicts, but also why. Our modality ablation study further demonstrates that each data modality contributes a non-redundant signal, with time-series data offering the largest individual performance gain and ICD codes and radiography providing complementary improvements, highlighting the necessity of structured multi-modal integration rather than reliance on a single dominant feature modality.

Our work, however, has limitations. While we account for competing risks and multi-modality, the model does not explicitly incorporate physician actions (treatments or interventions), which can impact both event timing and modality relevance. Furthermore, although we demonstrate generalisability across a diverse set of five real-world datasets, external validation on datasets from other health systems will be essential before deployment. Future work will explore subgroup-specific attention maps to audit the model for potential biases and further stratify interpretability results by demographics such as age, sex, and race.

In summary, our model advances the field of time-to-event prediction by introducing a novel framework for integrating multi-modal EHR data using dynamic graph learning with hierarchical interpretability. Our model yields robust improvements in discrimination, calibration, and transparency over existing methods and opens new directions for clinically grounded, interpretable machine learning in healthcare.

A Code and Data

A.1 Data Description

The pre-processing pipeline for MIMIC-IV was based on previously published workflows, and eICU was based in part on work done by [26, 30]. We used the imputation as suggested by the pipeline.

For the time-series variables, we use forward filling as clinicians in practice would only consider the last recorded measurement. If the first set of measurements is missing for some time-varying features, instead of dropping those features or patients, we backward-fill from the closest measurement in the future. The time-series features were resampled to 1-hour intervals. For the ICU datasets, we considered only observations collected up to 24 hours before the registered outcome. For MC-MED, since it is an ED dataset, the entirety of the patient cohort is within 24 hours of stay within the emergency department, and we include all of this information before the event time itself. For PBC2, we resampled the data into a monthly timescale. Patient admissions were randomly split into train, validation and test sets (8:1:1). Details of the features included can be found in Supplementary Tables 1, 2, and 3.

For eICU, MIMIC-IV, and MC-MED, the data contains de-identified patient electronic health records data, which can only be obtained after the ethical review of the proposed analysis on the PhysioNet page. Some certification of training modules is also required for access. We have cited the sources for the datasets in the text accordingly under Data. Consent for data use has been obtained by the providers, de-identification and licensing are in line with HIPAA requirements and compatible with the research conducted, which has passed ethical review and certification for data access.

The most relevant feature distributions for eICU, MIMIC-IV, and MC-MED can be found in Table 6. The list of all features are in tables 8, 7 9, and 10.

A.2 Code, Benchmark Models, and Training Details

Sample data and code implementations can be found here: https://anonymous.4open.science/r/Multi-Modal-Graph-A1BC/README.md. The repository includes implementations for all baseline and proposed models, with a provided `requirements.txt` specifying package dependencies and versions.

We evaluate our model against two groups of baselines for both single-risk and competing-risk survival analysis tasks. The first group comprises classical and deep survival models: Cox Proportional Hazards (Cox PH), DeepSurv, DeepHit, Dynamic-DeepHit, and DySurv. The second group includes graph-based survival models: GCN, GAT, GraphSAGE, TodyNet, DynaGraph, MedGNN, and MM-STGNN. For MM-STGNN, we replace the original imaging modality with embeddings from radiographic reports to match our input setup and adapt it with a competing-risk loss function. The same adaptation is made for MedGNN. For TodyNet and DynaGraph, we retain only the time-series and/or static modalities as supported by each dataset.

All models were implemented in PyTorch and trained on a single NVIDIA V100 GPU with 50 GB RAM. Data loading and training were fully reproducible using fixed seeds: 42, 1992, 1709, 250, and 213. Validation performance was used for early stopping and final model selection.

The Adam optimizer was used throughout, with a default learning rate of 0.001 unless otherwise noted. For benchmark models, we used a batch size of 32 for eICU, MIMIC-IV, and MC-MED datasets. For our model, we searched over {8, 13, 32, 64}, with 32 found optimal. Epochs were set to 10 for eICU, 11 for MIMIC-IV, and 10 for MC-MED. Each full run on MIMIC-IV takes approximately 23 minutes.

We performed grid search on MIMIC-IV to tune hyperparameters. Details of hyperparameter ranges and selected values are given below in Table 11.

B Metrics

In this section, we switch sample notation from superscripts to subscripts, i.e., $\mathbf{X}^i$ becomes $\mathbf{X}_i$ for patient i. Since our model predicts full event-time distributions under competing risks and right-censoring, rather than single-time binary labels, classical classification metrics are insufficient. We rely on metrics designed specifically for survival analysis, including both ranking-based and calibration-based scores.

The most widely used metric in survival modelling is the *concordance index* (C_{ind}), which estimates the probability that, for a randomly selected comparable pair of patients, the model correctly ranks their relative event times [16]. While straightforward under proportional hazards assumptions, where risk rankings are time-invariant, we instead use the time-dependent extension $C^{\text{td}}_{\text{ind}}$ that accommodates time-varying survival predictions [1]. Formally:

$$C^{\text{td}}_{\text{ind}} = \mathbb{P}\left\{\hat{F}(t_i \mid \mathbf{X}_i) > \hat{F}(t_i \mid \mathbf{X}_j) \mid t_i < t_j, \epsilon_i \neq \varnothing\right\} \tag{6}$$

where $\hat{F}(t \mid \mathbf{X})$ denotes the predicted cumulative incidence (for any event) at time t. Only uncensored events ($\epsilon_i \neq \varnothing$) are considered for the evaluation.

In the competing risks setting, each patient may experience one of multiple mutually exclusive event types. Therefore, concordance must be evaluated separately for each event type to assess how well the model ranks time-to-event predictions per cause. The cause-specific time-dependent concordance index for event ϵ is defined as:

$$C^{\text{td}}_{\text{ind},\epsilon} = \mathbb{P}\left\{\hat{F}_\epsilon(t_i \mid \mathbf{X}_i) > \hat{F}_\epsilon(t_i \mid \mathbf{X}_j) \mid t_i < t_j, \epsilon_i = \epsilon\right\} \tag{7}$$

where $\hat{F}_\epsilon(t \mid \mathbf{X})$ is the predicted cumulative incidence function (CIF) for event type ϵ. Only pairs where patient i experienced event ϵ before patient j are considered.

We report both per-cause $C^{\text{td}}_{\text{ind},\epsilon}$ and the macro-averaged concordance across all non-censoring events:

$$\bar{C}^{\text{td}}_{\text{ind}} = \frac{1}{E}\sum_{\epsilon=1}^{E} C^{\text{td}}_{\text{ind},\epsilon} \tag{8}$$

where E is the total number of event types.

The *Brier Score* measures the squared difference between the predicted and true survival probabilities at a given timepoint, similar to mean squared error, and ranges from 0 (best) to 1 (worst) [6, 15]. To adjust for censoring, we use the inverse probability of censoring

Table 6: Summary of demographics and clinical variables across three datasets: eICU, MIMIC-IV, and MC-MED. MIMIC-IV was used exclusively as an external validation set.

Attributes	eICU (N = 82,155)	MIMIC-IV (N = 71,935)	MC-MED (N = 23,128)
Age (mean ± SD)	67.2 ± 12.4	74.1 ± 13.4	58.3 ± 18.9
Sex (male)	64.5%	57.1%	46.7%
Ethnicity (Caucasian)	77.3%	71.1%	40.2%
Ethnicity (African American)	10.6%	8.0%	6.5%
Ethnicity (Hispanic)	3.7%	3.1%	18.6%
Ethnicity (Asian)	1.6%	2.0%	13.3%
Lactate (mmol/L)	2.5 ± 2.3	2.0 ± 1.5	2.4 ± 2.1
SBP (mmHg)	120.0 ± 16.3	126.3 ± 18.8	129.2 ± 17.2
Glucose (mg/dL)	147.3 ± 56.7	136.5 ± 49.3	138.4 ± 51.0
WBC ($\times 10^9$/L)	15.1 ± 9.3	10.6 ± 7.4	9.9 ± 6.1
RDW (%)	15.0 ± 2.0	14.4 ± 2.1	14.1 ± 2.0
Urea Nitrogen (mg/dL)	22.8 ± 13.4	22.8 ± 17.0	24.1 ± 16.8
Bicarbonate (mmol/L)	24.8 ± 4.4	23.3 ± 3.1	23.9 ± 4.2
Mortality	12.0%	9.7%	2.3% (ICU admission)

Table 7: Features extracted from the MIMIC-IV database. The features include demographic data collected for all patients, ICU unit-specific information like the type of unit, hospital information, vital signs, and biochemical measurements.

Static Variables			
Feature	*Type*	*Feature*	*Type*
Sex	binary	Admission Type	categorical
Age	integer	Insurance	categorical
ICU Type	categorical	Ethnicity	categorical
Time-series Variables			
Feature	*Type*	*Feature*	*Type*
Anion Gap	continuous	WBC	continuous
Weight	continuous	Temperature	continuous
SBP	continuous	DBP	continuous
Sodium	continuous	Respiratory Rate	continuous
RBC	continuous	Prothrombin Time PT	continuous
Prothrombin Time INR	continuous	Potassium	continuous
Platelets	categorical	Phosphorous	continuous
Phosphate	continuous	Partial Thromboplastin Time	continuous
Oxygen Saturation	continuous	MCGC	continuous
Magnesium	continuous	Hemoglobin	continuous
Hematocrit	continuous	Heart Rate	continuous
Glucose	continuous	Chloride	continuous
Creatinine	continuous	Calcium	continuous
BUN	continuous	Bicarbonate	continuous
Vent	binary	Vaso	binary
Adenosine	binary	Dobutamine	binary
Dopamine	binary	Epinephrine	binary
Isuprel	binary	Milrinone	binary
Norepinephrine	binary	Phenylepinephrine	binary
Vasopressin	binary	Colloid	binary
Crystalloid	binary	Intervention Duration	binary

weighted (IPCW) Brier Score:

$$\mathrm{BS}_{\mathrm{IPCW}}(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\hat{S}(t \mid \mathbf{X}_i)^2 \mathbb{1}\{t_i \le t, \epsilon_i \neq \varnothing\}}{\hat{G}(t_i^-)} + \frac{(1-\hat{S}(t \mid \mathbf{X}_i))^2 \mathbb{1}\{t_i > t\}}{\hat{G}(t)}\right] \tag{9}$$

where $\hat{S}(t \mid \mathbf{X}_i)$ is the predicted survival probability, and $\hat{G}(t)$ is the Kaplan–Meier estimate of the censoring distribution.

Table 8: Features extracted from the eICU database. The features include demographic data collected for all patients, ICU unit-specific information like type and number of beds, hospital information like regional location and teaching status, vital signs including respiratory rate and blood pressure, and biochemical measurements including troponin and levels of potassium and protein in the blood.

Feature	Type	Feature	Type
Sex	binary	Unit Stay Type	categorical
Age	integer	Num Beds Category	categorical
Height	continuous	Region	categorical
Weight	continuous	Teaching Status	binary
Ethnicity	categorical	Physician Speciality	categorical
Unit Type	categorical	Unit Type	categorical
Unit Admit Source	categorical	Mechanical Ventilation	binary
Unit Visit Number	categorical		
Time-series Variables			
Feature	*Type*	*Feature*	*Type*
		Base Excess	continuous
-basos	continuous	FiO2	continuous
-eos	continuous	HCO3	continuous
-monos	continuous	Hct	continuous
-polys	continuous	Hgb	continuous
ALT	continuous	MCH	continuous
AST	continuous	MCHC	continuous
BUN	continuous	MCV	continuous
O2 Sat (%)	continuous	MPV	continuous
PT-INR	continuous	PT	continuous
RBC	continuous	PTT	continuous
RDW	continuous	WBC	continuous
Alkaline ph.	continuous	Albumin	continuous
Bedside Glucose	continuous	Anion Gap	continuous
Calcium	continuous	Bicarbonate	continuous
Creatinine	continuous	Glucose	continuous
Lactate	continuous	Magnesium	continuous
pH	continuous	paCO2	continuous
paO2	continuous	Phosphate	continuous
Platelets	continuous	Potassium	continuous
Sodium	continuous	Bilirubin	continuous
Protein	continuous	Troponin - I	continuous
Urinary s. Gravity	continuous	mean BP	continuous
SBP	continuous	DBP	continuous

The *Integrated Brier Score (IBS)* averages the Brier Score over time:

$$\text{IBS} = \frac{1}{t_{\max}} \int_0^{t_{\max}} \text{BS}_{\text{IPCW}}(t)\, dt \tag{10}$$

where $t_{\max}$ is the maximum observed time.

Finally, we assess probabilistic calibration using the IPCW-adjusted *binomial log-likelihood* (BLL), which measures the accuracy of predicted survival probabilities as probabilistic forecasts:

$$\text{BLL}(t) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\log(1 - \hat{S}(t \mid \mathbf{X}_i)) \mathbb{1}\{t_i \le t, \epsilon_i \neq \varnothing\}}{\hat{G}(t_i)} + \frac{\log(\hat{S}(t \mid \mathbf{X}_i)) \mathbb{1}\{t_i > t\}}{\hat{G}(t)} \right] \tag{11}$$

Table 9: Features extracted from the MC-MED database. The features include demographic data, triage variables, vital signs, biochemical measurements, as well as ICD diagnosis codes and radiography embeddings.

Static Variables			
Feature	*Type*	*Feature*	*Type*
Sex	binary	Ethnicity	categorical
Age	integer	Triage Acuity Score	ordinal
Time-series Variables			
Feature	*Type*	*Feature*	*Type*
HR	continuous	RR	continuous
SBP	continuous	DBP	continuous
SpO2	continuous	Temp	continuous
Pain Score	continuous	Perfusion Index	continuous
1min HRV	continuous	5min HRV	continuous
BUN	continuous	Creatinine	continuous
Sodium	continuous	Potassium	continuous
Chloride	continuous	Bicarbonate	continuous
Calcium	continuous	Corrected Calcium	continuous
Albumin	continuous	Globulin	continuous
Total Protein	continuous	Bilirubin (Total)	continuous
AST (SGOT)	continuous	ALT (SGPT)	continuous
Alkaline Phosphatase	continuous	Anion Gap	continuous
Hemoglobin	continuous	Hematocrit	continuous
MCV	continuous	MCH	continuous
MCHC	continuous	RBC	continuous
WBC	continuous	Platelet Count	continuous
EGFR (no race)	continuous	Glucose	continuous
Triage Vital Signs (HR, RR, SBP, DBP, Temp, SpO2)	continuous	–	–
ICD Codes	multi-hot vectors (500 most frequent codes)		
Radiography Embeddings	Dense vector representations (768-dim embeddings)		

Its integrated version, the *Integrated Binomial Log-Likelihood (IBLL)*, is:

$$\text{IBLL} = \frac{1}{t_{\max}} \int_0^{t_{\max}} \text{BLL}(t)\, dt \tag{12}$$

All metrics are computed on the held-out test set. Time integrals are approximated using numerical integration over 100 uniformly spaced timepoints, consistent with prior work [16]. Together, $C^{\text{td}}_{\text{ind}}$, IBS, and IBLL provide complementary perspectives on model performance, evaluating both temporal discrimination and calibration under competing risks.

C Interpretability

To enable transparency in predictions, we integrate a pseudo-attention mechanism during graph construction. Each graph at time slice s is associated with a learnable interpretability matrix $I_s \in \mathbb{R}^{d \times d}$, which captures the relative contribution of nodes and edges to the loss. These matrices are initialised uniformly and updated via backpropagation based on their influence on the model's output.

For a node v, we define its importance score by combining its individual gradient saliency and the average importance of its connected edges:

$$I_v = \alpha \cdot \|\nabla_{h_v} L\| + (1 - \alpha) \cdot \frac{1}{|\mathcal{N}(v)|} \sum_{u \in \mathcal{N}(v)} \|\nabla_{e_{v,u}} L\|, \tag{13}$$

where $\alpha \in [0, 1]$ balances node and edge contributions, h_v is the embedding of node v, $e_{v,u}$ is the edge from v to u, and L is the task loss. This formulation allows us to attribute model behaviour to individual clinical variables and their temporal interactions.

These interpretability matrices are computed independently for each modality (e.g., vitals, demographics, ICD codes, radiography) and later fused, preserving hierarchical structure. The fused matrix informs both graph construction and attention-weighted pooling, offering a consistent mechanism for identifying influential features over time.

In contrast to traditional post-hoc methods, our integrated approach ensures interpretability is aligned with the model's learned representations and dynamically adapts across training epochs. This is critical in healthcare, where understanding temporal feature salience can support trust and accountability in clinical decision-making.

D Convergence of the Hierarchical Interpretability Matrix

Definitions and Assumptions

Let $G_t = (V, E_t, I_t)$ denote the graph at epoch t, where:

- V is the set of nodes (e.g., patient-specific features across modalities).

Table 10: Summary of features used in the PBC2 and SUPPORT datasets. PBC2 includes liver disease-specific biomarkers and clinical indicators, while SUPPORT comprises general clinical, physiological, and comorbidity-related variables for hospitalized patients.

Category	PBC2		SUPPORT	
	Feature	*Type*	*Feature*	*Type*
Demographics	Age	continuous	Age	continuous
	Sex	binary	Sex	binary
			Race	categorical
			DNR Status (Day 1)	binary
Clinical	Edema	categorical	Coma Score	ordinal
	Ascites	binary	Cancer	binary
	Hepatomegaly	binary	CHF	binary
	Spiders (vascular lesions)	binary	Cirrhosis	binary
Biomarkers / Labs	Albumin	continuous	Serum Albumin	continuous
	Bilirubin	continuous	Serum Sodium	continuous
	ALP	continuous	Serum Potassium	continuous
	AST (SGOT)	continuous	Serum Creatinine	continuous
	Prothrombin Time	continuous	Hematocrit	continuous
	Platelets	continuous	WBC Count	continuous
			PaO_2	continuous
Vitals			Systolic BP	continuous
			Heart Rate	continuous
			Respiratory Rate	continuous
			Temperature	continuous
			Glasgow Coma Scale	continuous

Table 11: Hyperparameter search ranges for all models. Bold values indicate optimal parameters found via grid search on the validation set. All models were trained using the Adam optimiser.

Model	Batch Size	Learning Rate	Dropout	Graph Layers	MLP Layers
Our Model	8, 16, **32**, 64	0.01, 0.001, **0.0001**	0.5, **0.7**, 0.9	**2**, 3, 4	**2**, 4, 6
DeepSurv	32, **64**, 128	0.01, 0.001, **0.0001**	0.5, **0.7**, 0.9	–	2, **4**, 6
DeepHit	32, 64, **128**	0.01, 0.001, **0.0001**	0.5, **0.7**, 0.9	–	2, **4**, 6
Dynamic-DeepHit	32, 64, **128**	0.01, 0.001, **0.0001**	0.5, **0.7**, 0.9	–	2, **4**, 6
DySurv	**32**, 64, 128	**0.001**, 0.0001	0.5, **0.7**, 0.9	–	2, **4**, 6
GCN	32, 64, **128**	0.01, 0.001, **0.0001**	**0.5**, 0.7, 0.9	2, **3**, 4	2, **4**, 6
GAT	32, 64, **128**	0.01, 0.001, **0.0001**	**0.5**, 0.7, 0.9	2, **3**, 4	2, **4**, 6
GraphSAGE	32, 64, **128**	**0.001**, 0.0001	0.5, **0.7**, 0.9	2, **3**, 4	2, **4**, 6
TodyNet	32, **64**, 128	**0.001**, 0.0001	0.5, **0.7**, 0.9	**2**, 3, 4	2, **4**, 6
DynaGraph	32, 64, **128**	**0.0001**, 0.001	0.5, **0.7**, 0.9	**2**, 3, 4	2, **4**, 6
MedGNN	32, 64, **128**	**0.0001**, 0.001	0.5, **0.7**, 0.9	2, 3, **4**	2, **4**, 6
MM-STGNN	**32**, 64, 128	**0.0001**, 0.001	0.5, **0.7**, 0.9	**2**, 3, 4	2, **4**, 6

- E_t is the set of edges determined by temporal and modality-level relationships.
- I_t is the fused interpretability matrix at epoch t, derived from hierarchical attention across modalities and time.

Let $I_t = f(I_t^{\text{dyn}}, I_t^{\text{static}}, I_t^{\text{ICD}}, I_t^{\text{rad}})$ be the fused matrix at epoch t, computed from the attention matrices associated with each modality through a learnable aggregation function $f(\cdot)$.

Convergence Definition: The hierarchical interpretability matrix converges if:

$$\lim_{t\to\infty} \|I_t - I_{t-1}\|_F = 0,$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Assumptions:

(1) Each modality-specific attention matrix $I_t^{(m)}$ is optimized via backpropagation with a differentiable loss function $\mathcal{L}$ and follows gradient descent updates.
(2) The fused interpretability matrix I_t is differentiable with respect to all $I_t^{(m)}$, and the fusion function $f(\cdot)$ is smooth (e.g., weighted sum or attention-based).
(3) The learning rate η_t satisfies $\eta_t \to 0$, $\sum_{t=1}^{\infty} \eta_t = \infty$, and $\sum_{t=1}^{\infty} \eta_t^2 < \infty$.

Proof of Convergence

We begin with the gradient update rule for each modality-specific interpretability matrix:

$$I_t^{(m)} = I_{t-1}^{(m)} - \eta_t \nabla_{I^{(m)}} \mathcal{L}(I_{t-1}^{(m)}).$$

Assuming $f(\cdot)$ is differentiable and linear or Lipschitz-continuous, the fused matrix evolves as:

$$I_t = f(I_t^{\text{dyn}}, I_t^{\text{static}}, I_t^{\text{ICD}}, I_t^{\text{rad}}).$$

Applying the multivariate Taylor expansion of $\mathcal{L}$ at $I_{t-1}^{(m)}$, for each modality:

$$\mathcal{L}(I_t^{(m)}) \approx \mathcal{L}(I_{t-1}^{(m)}) - \eta_t \|\nabla_{I^{(m)}} \mathcal{L}(I_{t-1}^{(m)})\|_F^2 + \frac{1}{2}\eta_t^2 \|H^{(m)}\|_F \|\nabla_{I^{(m)}} \mathcal{L}(I_{t-1}^{(m)})\|_F^2,$$

where $H^{(m)}$ is the Hessian of $\mathcal{L}$ with respect to $I^{(m)}$. Assuming bounded Hessians and the learning rate conditions, the second-order term vanishes faster, so:

$$\lim_{t\to\infty} \|\nabla_{I^{(m)}} \mathcal{L}(I_{t-1}^{(m)})\|_F = 0.$$

Since each modality-specific matrix $I_t^{(m)}$ converges and $f(\cdot)$ is continuous, their composition $I_t = f(I_t^{(m)})$ also converges. Hence:

$$\lim_{t\to\infty} \|I_t - I_{t-1}\|_F = 0.$$

Conclusion: Under standard assumptions on smoothness of the loss function and learning rates, the fused interpretability matrix constructed through hierarchical attention converges. This ensures the interpretability weights used in the model stabilise, making the model's explanations consistent and trustworthy over training epochs.

E Graph Learning with GIN

Continuing from the Methods, we first construct a fused graph representation for each sample i by concatenating the adjacency matrix $A^{(i)}$ with the corresponding modality-level interpretability matrix $I^{(i)}$:

$$G^{(i)} = \left(A^{(i)} \parallel I^{(i)}\right) \tag{14}$$

where $\parallel$ denotes concatenation along the feature dimension. This fusion integrates both structural and attention-based relational information across modalities and time.

To extract higher-order node representations from these fused graphs, we employ a multi-layer Graph Isomorphism Network (GIN), which has been shown to be effective at capturing complex node interactions and topological patterns [27]. The GIN layers iteratively update node embeddings through neighborhood aggregation, capturing temporal and multimodal dependencies in the graph structure. Formally, for a k-layer GIN, the node embeddings are updated as:

$$h_v^{(k)} = \text{MLP}^{(k)}\left((1+\epsilon^{(k)}) \cdot h_v^{(k-1)} + \sum_{u\in\mathcal{N}(v)} h_u^{(k-1)}\right) \tag{15}$$

where $h_v^{(k)}$ is the embedding of node v at layer k, $\mathcal{N}(v)$ denotes the neighbors of node v, $\epsilon^{(k)}$ is a learnable scalar, and $\text{MLP}^{(k)}$ is a multi-layer perceptron specific to the k-th layer.

The final node embeddings $h^{(K)}$ from the last GIN layer are then passed into the temporal pooling module described in Section F, which compresses these embeddings hierarchically while preserving essential temporal dynamics. This pipeline allows the model to produce rich, temporally-aware, and topologically-grounded representations for downstream competing risks survival prediction.

F Temporal Pooling

After obtaining spatio-temporal embeddings from the hierarchical GIN module, we apply a learnable temporal pooling mechanism to compress the temporal graph representation while preserving salient temporal and relational structures across time. Instead of flattening the node-time embeddings, which risks losing meaningful temporal dependencies, we adopt a hierarchical 2D convolution-based pooling strategy inspired by [22]. This approach enables the model to learn soft assignments of temporal nodes into clusters, producing compact graph representations suitable for downstream survival prediction.

Formally, given an input tensor $X^l \in \mathbb{R}^{N^l \times d}$ at the l-th pooling layer (where N^l is the number of temporal graph nodes and d the embedding dimension), we apply a 2D convolution over the temporal axis (treated as channels), producing an output $X^{l+1} \in \mathbb{R}^{N^{l+1}\times d}$:

$$X^{l+1} = \sum_{j=0}^{N^l-1} W(N^{l+1}, j) \star X^l + b(N^{l+1})$$

where W are learnable convolution weights, b is a bias term, and $\star$ denotes cross-correlation. This operation yields new cluster-level node embeddings while reducing the node count from N^l to N^{l+1}.

To propagate structural information through the hierarchy, we reconstruct lower-level adjacency matrices based on the learned soft clustering. Let $W^l \in \mathbb{R}^{N^{l+1}\times N^l \times 1 \times k}$ be the 2D convolutional weights reshaped into a matrix $M^l \in \mathbb{R}^{N^{l+1}\times N^l}$ using a learnable vector $V^l \in \mathbb{R}^{1\times k}$ such that $M^l = W^l \cdot V^l$. The coarsened adjacency matrix is then computed as:

$$A^{l+1} = M^l A^l {M^l}^\top \in \mathbb{R}^{N^{l+1}\times N^{l+1}}$$

This allows the model to maintain a graph structure at each level of abstraction, where A^{l+1} captures the weighted connectivity between temporal clusters. Crucially, both node and edge representations evolve jointly and hierarchically, enabling end-to-end optimisation of both temporal resolution and relational importance. The final pooled representations X^{l+1} and A^{l+1} are passed to cause-specific multi-layer perceptions for competing risk survival prediction.

G Full Model Evaluation Results

We provide a comprehensive assessment of all evaluated models across three standard metrics in survival analysis: cause-specific time-dependent concordance index (C-index), Integrated Brier Score (IBS), and Integrated Binomial Log-Likelihood (IBLL). These metrics respectively measure the discriminative performance, calibration, and probabilistic accuracy of survival models.

Our proposed model achieves consistently superior performance across all five datasets (MIMIC-IV, eICU, PBC2, MC-MED, SUPPORT) when compared to both traditional survival methods (e.g.,

Cox PH), temporal deep learning models (e.g., DeepHit, Dynamic-DeepHit, DySurv), and graph-based approaches (e.g., GCN, GAT, GraphSAGE, MM-STGNN, MedGNN, DynaGraph, TodyNet). This robustness across diverse settings underscores the effectiveness of our model's dynamic, multi-modal graph architecture with hierarchical attention and temporal pooling.

Figure 2a shows a spider plot of concordance scores across datasets, clearly illustrating our model's consistent edge in discriminative power. Supplementary Figures 2b and 2c provide the corresponding plots for IBS and IBLL metrics, revealing strong calibration and generalisation performance across datasets as well.

In all metrics, performance gains are especially pronounced on complex, multi-modal datasets like MIMIC-IV and MC-MED, where our model benefits from its ability to jointly represent static, temporal, diagnostic, and radiographic modalities. These improvements are also evident in smaller datasets like PBC2 and SUPPORT, demonstrating the model's adaptability to varying cohort sizes and clinical settings.

H Calibration Plots

Results of calibrations can be seen in Figure 3.

References

[1] Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. 2005. A time-dependent discrimination index for survival data. *Statistics in medicine* 24, 24 (2005), 3927–3944.

[2] Peter C Austin, Douglas S Lee, and Jason P Fine. 2016. Introduction to the analysis of survival data in the presence of competing risks. *Circulation* 133, 6 (2016), 601–609.

[3] Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. 2020. Adaptive graph convolutional recurrent network for traffic forecasting. *Advances in neural information processing systems* 33 (2020), 17804–17815.

[4] Defu Cao, Yujing Wang, Juanyong Duan, Ce Zhang, Xia Zhu, Congrui Huang, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, et al. 2020. Spectral temporal graph neural network for multivariate time-series forecasting. *Advances in neural information processing systems* 33 (2020), 17766–17778.

[5] Wei Fan, Jingru Fei, Dingyu Guo, Kun Yi, Xiaozhuang Song, Haolong Xiang, Hangting Ye, and Min Li. 2025. MedGNN: Towards Multi-resolution Spatiotemporal Graph Learning for Medical Time Series Classification. *arXiv preprint arXiv:2502.04515* (2025).

[6] Stephane Fotso. 2018. Deep neural networks for survival analysis based on a multi-task framework. *arXiv preprint arXiv:1801.05512* (2018).

[7] Rebecca Giddings, Anabel Joseph, Thomas Callender, Sam M Janes, Mihaela Van der Schaar, Jessica Sheringham, and Neal Navani. 2024. Factors influencing clinician and patient interaction with machine learning-based risk prediction models: a systematic review. *The Lancet Digital Health* 6, 2 (2024), e131–e144.

[8] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems* 30 (2017).

[9] Na Hong, Chun Liu, Jianwei Gao, Lin Han, Fengxiang Chang, Mengchun Gong, and Longxiang Su. 2022. State of the art of machine learning–enabled clinical decision support in intensive care units: literature review. *JMIR medical informatics* 10, 3 (2022), e28781.

[10] Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P Lungren. 2020. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ digital medicine* 3, 1 (2020), 136.

[11] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. 2018. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC medical research methodology* 18, 1 (2018), 1–12.

[12] Boris Khodorkovsky, Elias Youssef, Frosso Adamakos, Tiffany Cina, Amanda Falco, Lauren LaMura, Anthony Marion, Samuel Nathan, and Barry Hahn. 2018. Does initial temperature in the emergency department predict outcomes in patients admitted for sepsis? *The Journal of Emergency Medicine* 55, 3 (2018), 372–377.

[13] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[14] Anders Kasper Bruun Kristensen, Jon Gitz Holler, Søren Mikkelsen, Jesper Hallas, and Annmarie Lassen. 2015. Systolic blood pressure and short-term mortality in the emergency department and prehospital setting: a hospital-based cohort study. *Critical Care* 19 (2015), 1–8.

[15] Håvard Kvamme and Ørnulf Borgan. 2019. The brier score under administrative censoring: Problems and solutions. *arXiv preprint arXiv:1912.08581* (2019).

[16] Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. 2019. Time-to-event prediction with neural networks and Cox regression. *Journal of machine learning research* 20, 129 (2019), 1–30.

[17] Changhee Lee, Jinsung Yoon, and Mihaela Van Der Schaar. 2019. Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering* 67, 1 (2019), 122–133.

[18] Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. 2018. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[19] Jianing Li, Thomas H Scheike, and Mei-Jie Zhang. 2015. Checking Fine and Gray subdistribution hazards model with cumulative sums of residuals. *Lifetime data analysis* 21, 2 (2015), 197–217.

[20] Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2022. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838* (2022).

[21] Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2023. A comparative study of pretrained language models for long clinical text. *Journal of the American Medical Informatics Association* 30, 2 (2023), 340–347.

[22] Huaiyuan Liu, Xianzhang Liu, Donghua Yang, Zhiyu Liang, Hongzhi Wang, Yong Cui, and Jun Gu. 2023. TodyNet: Temporal Dynamic Graph Neural Network for Multivariate Time Series Classification. *arXiv preprint arXiv:2304.05078* (2023).

[23] Huaiyuan Liu, Donghua Yang, Xianzhang Liu, Xinglei Chen, Zhiyu Liang, Hongzhi Wang, Yong Cui, and Jun Gu. 2024. Todynet: temporal dynamic graph neural network for multivariate time series classification. *Information Sciences* 677 (2024), 120914.

[24] Torben Martinussen. 2022. Causality and the Cox regression model. *Annual Review of Statistics and Its Application* 9, 1 (2022), 249–259.

[25] Munib Mesinovic, Soheila Molaei, Peter Watkinson, and Tingting Zhu. 2025. DynaGraph: Interpretable Multi-Label Prediction from EHRs via Dynamic Graph Learning and Contrastive Augmentation. *arXiv preprint arXiv:2503.22257* (2025).

[26] Munib Mesinovic, Peter Watkinson, and Tingting Zhu. 2024. DySurv: dynamic deep learning model for survival analysis with conditional variational inference. *Journal of the American Medical Informatics Association* (2024), ocae271.

[27] Ciyuan Peng, Jiayuan He, and Feng Xia. 2024. Learning on multimodal graphs: A survey. *arXiv preprint arXiv:2402.05322* (2024).

[28] Jiajun Qiu, Yao Hu, Li Li, Abdullah Mesut Erzurumluoglu, Ingrid Braenne, Charles Whitehurst, Jochen Schmitz, Jatin Arora, Boris Alexander Bartholdy, Shrey Gandhi, et al. 2025. Deep representation learning for clustering longitudinal survival data from electronic health records. *Nature Communications* 16, 1 (2025), 2534.

[29] Md Mahmudur Rahman, Koji Matsuo, Shinya Matsuzaki, and Sanjay Purushotham. 2021. Deeppseudo: Pseudo value based deep learning models for competing risk analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 479–487.

[30] Emma Rocheteau, Pietro Liò, and Stephanie Hyland. 2021. Temporal pointwise convolutional networks for length of stay prediction in the intensive care unit. In *Proceedings of the conference on health, inference, and learning*. 58–68.

[31] Jörg Schilcher, Alva Nilsson, Oliver Andlid, and Anders Eklund. 2024. Fusion of electronic health records and radiographic images for a multimodal deep learning prediction model of atypical femur fractures. *Computers in Biology and Medicine* 168 (2024), 107704.

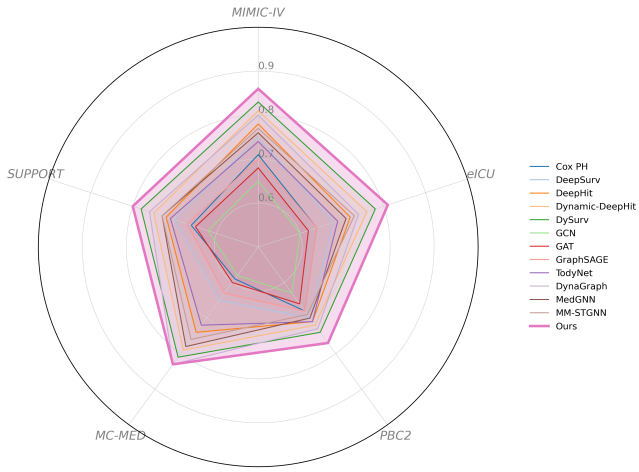
[32] Alice S Tang, Sarah R Woldemariam, Silvia Miramontes, Beau Norgeot, Tomiko T Oskotsky, and Marina Sirota. 2024. Harnessing EHR data for health research. *Nature Medicine* 30, 7 (2024), 1847–1855.

[33] Siyi Tang, Amara Tariq, Jared A Dunnmon, Umesh Sharma, Praneetha Elugunti, Daniel L Rubin, Bhavik N Patel, and Imon Banerjee. 2023. Predicting 30-day all-cause hospital readmission using multimodal spatiotemporal graph neural networks. *IEEE Journal of Biomedical and Health Informatics* 27, 4 (2023), 2071–2082.

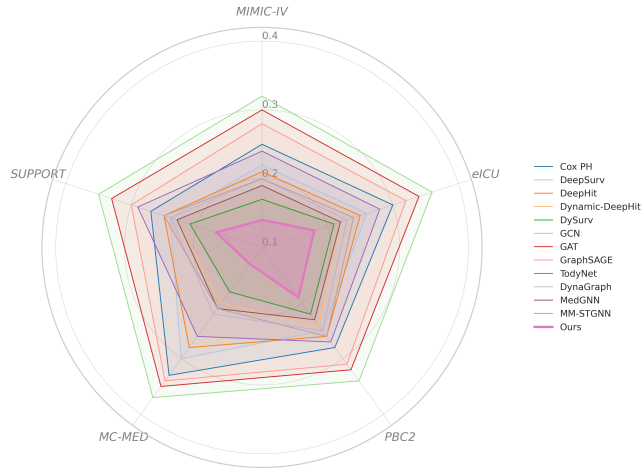
[34] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

[35] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. 2020. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 753–763.

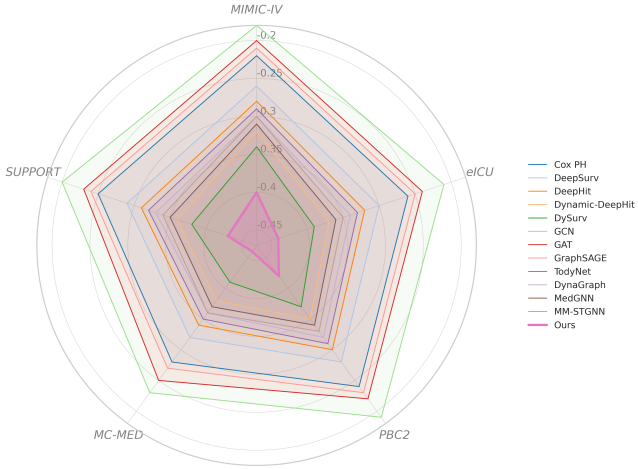
[36] Haoyan Xu, Ziheng Duan, Yueyang Wang, Jie Feng, Runjian Chen, Qianru Zhang, and Zhongbin Xu. 2021. Graph partitioning and graph neural network based hierarchical graph matching for graph similarity computation. *Neurocomputing* 439 (2021), 348–362.



(a) Cause-specific concordance index (higher is better).



(b) Integrated Brier Score (IBS; lower is better).



(c) Integrated binomial log-likelihood (IBLL; lower is better).

Figure 2: Performance of all models across five datasets (MIMIC-IV, eICU, PBC2, MC-MED, SUPPORT) measured using (a) cause-specific concordance index for the most common cause, (b) integrated Brier score (IBS), and (c) integrated binomial log-likelihood (IBLL). Our model consistently achieves the best performance across metrics and datasets.

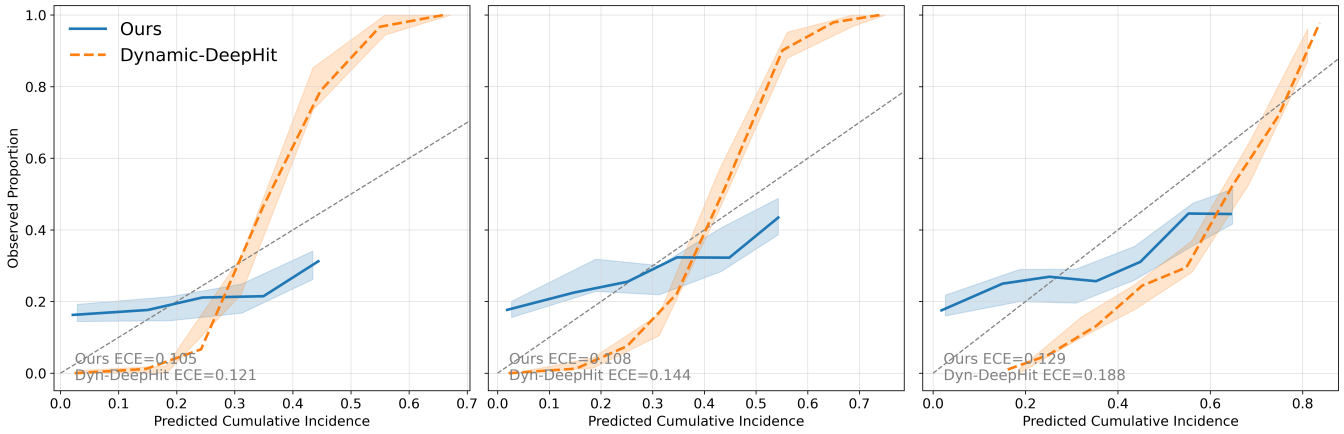
[37] Yishu Xue and Elizabeth D Schifano. 2017. Diagnostics for the Cox model. *Communications for Statistical Applications & Methods* 24, 6 (2017).

[38] Nor'azim Mohd Yunos, Rinaldo Bellomo, David McD Taylor, Simon Judkins, Fergus Kerr, Harvey Sutcliffe, Colin Hegarty, and Michael Bailey. 2017. Renal effects of an emergency department chloride-restrictive intravenous fluid strategy in patients admitted to hospital for more than 48 hours. *Emergency Medicine Australasia* 29, 6 (2017), 643–649.

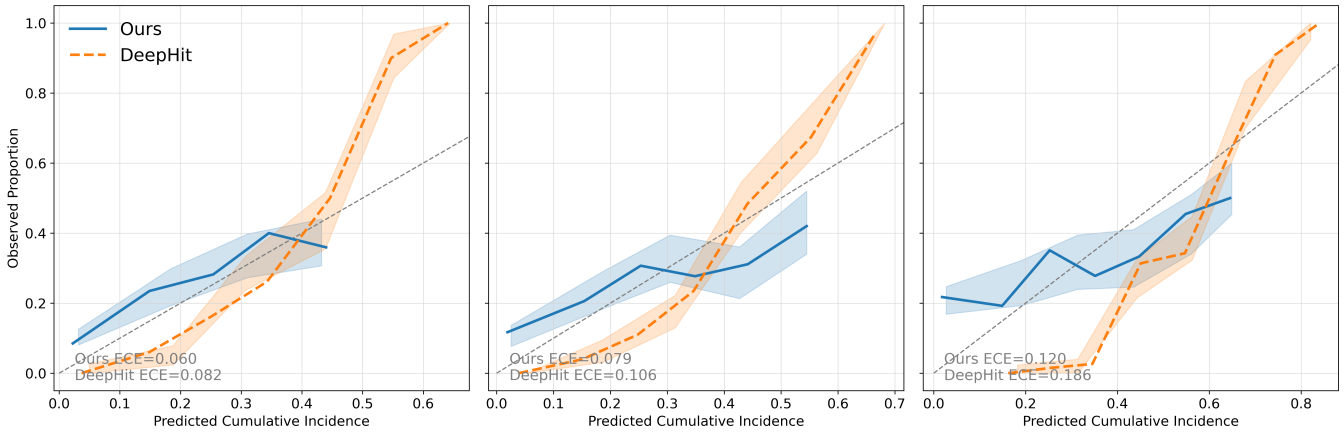
[39] Daochen Zha, Kwei-Herng Lai, Kaixiong Zhou, and Xia Hu. 2022. Towards similarity-aware time-series classification. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*. SIAM, 199–207.

[40] Zhongheng Zhang, Jaakko Reinikainen, Kazeem Adedayo Adeleke, Marcel E Pieterse, and Catharina GM Groothuis-Oudshoorn. 2018. Time-varying covariates and coefficients in Cox regression models. *Annals of translational medicine* 6, 7 (2018), 121.

[41] Zihan Zhang, Lei Shi, and Ding-Xuan Zhou. 2024. Classification with deep neural networks and logistic loss. *Journal of Machine Learning Research* 25, 125 (2024), 1–117.



(a) Calibration curves on MIMIC-IV dataset (ICU mortality at 48h, 96h, 144h).



(b) Calibration curves on eICU dataset (ICU mortality at 48h, 96h, 144h).

Figure 3: Calibration performance of our model versus Dynamic-DeepHit across multiple horizons and datasets. Panels (a) and (b) show results for MIMIC-IV and eICU, respectively. Each panel depicts calibration curves at 48h, 96h, and 144h prediction horizons. Curves are LOWESS-smoothed, and shaded regions indicate 80% bootstrap confidence intervals. Expected Calibration Error (ECE) is reported within each subplot. our model consistently yields better-aligned curves and lower ECE values, demonstrating improved calibration across ICU datasets.