Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.079891

ARTICLE

Interpretable Cox-Guided Risk Stratification for Specialized Expert Learning in Pan-Cancer Survival Prediction

Manal Mohammed AL-Tamimi[1,2,*], Siti Norul Huda Sheikh Abdullah[1,*], Mohammad Khatim Hasan[1], Mohammed Azmi Al-Betar[3,4], Maw Shin Sim[5] and Abdulrahman Mohammed AL-Tamimi[1]

[1]Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia
[2]Faculty of Administrative Science, Hadhramout University, AL-Mukalla, Yemen
[3]Artificial Intelligence Research Center (AIRC), College of Engineering and Information Technology, Ajman University, Ajman, United Arab Emirates
[4]Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman, Jordan
[5]Department of Pharmaceutical Life Sciences, Faculty of Pharmacy, Universiti Malaya, Kuala Lumpur, Malaysia

*Corresponding Authors: Manal Mohammed AL-Tamimi. Email: manal.altamimi@outlook.com; Siti Norul Huda Sheikh Abdullah. Email: snhsabdullah@ukm.edu.my

Received: 09 February 2026; Accepted: 15 April 2026; Published: 27 May 2026

ABSTRACT: Pan-cancer survival prediction remains a major challenge in personalized oncology due to profound tumor heterogeneity and the complexity of high-dimensional molecular data. Diverse risk profiles across cancer types and noisy, sparse features hinder deep learning models from capturing robust prognostic patterns. Prior pan-cancer studies predominantly focus on multimodal integration or unimodal gene expression analysis, leaving other informative modalities such as Copy Number Variation (CNV) and miRNA expression underexplored. We introduce a new formulation of mixture-of-experts (MoE) survival modeling that recasts expert assignment as a clinically interpretable risk-space decomposition problem. The proposed framework, CoxGuided-SE, constructs an ordered prognostic risk axis from a Cox proportional hazards model trained on clinical covariates, then uses quantile-based thresholds to deterministically stratify patients into risk-homogeneous groups and route them to specialized subnetworks. This design replaces black-box neural routing with a transparent, statistically principled alternative, preventing expert collapse without auxiliary regularization. Evaluated on 33 cancer types from TCGA across four modalities independently, namely clinical features, mRNA expression, miRNA expression, and CNV, CoxGuided-SE achieved substantial gains on high-dimensional genomic data, improving both discrimination and calibration. The most pronounced improvement occurred for CNV, widely considered one of the most challenging modalities for survival modeling, with a 24% increase in discrimination over the strongest baseline. The framework offers two-level interpretability: transparent patient routing based on clinical severity, and expert-specific feature importance within each subgroup. Feature-level analyses further confirm that experts learn complementary and non-redundant representations within each risk stratum. These results provide both a high-performing survival prediction framework and a principled explanation of when and why expert specialization is effective, advancing interpretable and trustworthy clinical AI for pan-cancer prognosis.

KEYWORDS: Survival prediction; mixture-of-experts; Cox proportional hazards; pan-cancer prognosis; deep learning; multi-omics; risk stratification; interpretable AI

1 Introduction

Accurate survival prognostication lies at the heart of clinical oncology, serving as a cornerstone for therapeutic decision making and patient management [1]. As the field advances toward precision medicine, and as cancer incidence and mortality continue to rise globally [2], the demand for individualized, data driven survival estimates has become increasingly urgent [3]. Deep learning models have become powerful tools for pan-cancer survival prediction, demonstrating strong performance on high-dimensional molecular and clinical data [4,5]. Unlike traditional Cox Proportional Hazards (CPH) models [6], which rely on linear assumptions, deep survival architectures such as SurvivalNet [7], Multi-Task Logistic Regression (MTLR) [8], and multimodal frameworks like MultiSurv [9], Fan et al. (2023) [10] can capture non-linear relationships and integrate diverse data sources. Recent advances have achieved state-of-the-art performance by leveraging large-scale molecular datasets across multiple cancer types [11–13]. However, these "one-size-fits-all" architectures, which process all patients through a single uniform network, present a fundamental limitation in the face of cancer heterogeneity. Cancer is not a single disease but a collection of highly heterogeneous conditions. Patients exhibit substantial diversity in clinical characteristics, molecular drivers, and disease trajectories, creating implicit domain shifts within pan-cancer cohorts [14–18]. A monolithic deep learning model, constrained to learn a single set of parameters, must approximate a compromised average function across these fundamentally different patient subpopulations. This limitation is particularly pronounced for high-dimensional molecular data such as mRNA expression, microRNA (miRNA) profiles, and copy number variations (CNV), where complex subgroup structure across thousands of features cannot be fully captured by a single global model. Consequently, uniform architectures can struggle with inconsistent performance, excelling in some patient subsets but failing in others [10,14,19].

Mixture-of-experts (MoE) architectures offer a promising framework for addressing this challenge by dividing the predictive task across specialized subnetworks. Each expert can focus on distinct patient subgroups, potentially improving generalization across heterogeneous cancer types. This paradigm has seen success in NLP and vision, and its ability to model variation makes it well-suited for pan-cancer prognosis. Growing interest in MoE architectures for oncology [20–24] has produced encouraging results; however, existing approaches exhibit several practical limitations that motivate further investigation. First, gating mechanisms typically rely on learned neural routing, whether through sparse top-k selection [23], soft learned weights [24], text-encoded projections [22], or fused-feature representations [20,21], where patient-to-expert assignment is not directly interpretable in terms of individual clinical characteristics. Second, in several frameworks, expert specialization is applied primarily at output layers [24] or fusion stages [20,21], which may limit the capacity of each expert to learn fully distinct representations. Third, many MoE survival models are evaluated on a limited number of cancer types (3–14) [20,21,23], and their behavior under large-scale pan-cancer heterogeneity remains less explored. Fourth, auxiliary load-balancing losses are often introduced to stabilize expert utilization [23,24], increasing training complexity. While learned gating can discover latent structure that may be difficult to specify a priori, it does not explicitly incorporate the observation that patients can be stratified into risk groups based on clinical prognostic variables whose contributions can be quantified from data.

This opacity contrasts with the long tradition of transparent risk stratification in clinical oncology, where scoring systems such as TNM staging [25] and the International Prognostic Index [26] assign patients to risk groups using explicit, human-readable criteria. However, such systems are manually designed and no unified scoring system covers all cancer types [27,28], motivating data-driven alternatives that preserve interpretability.

In this work, we introduce a new formulation of mixture-of-experts survival modelling, termed CoxGuided-SE (Cox-Guided Specialized Experts), that reframes expert assignment as a clinically interpretable risk-space decomposition problem rather than a latent pattern-discovery task. By leveraging Cox proportional hazards modelling to construct an interpretable and ordered prognostic risk axis, patients are deterministically stratified into risk-homogeneous groups and routed to specialized expert networks based on quantile thresholds of their clinical risk scores. This design conceptually mirrors how treatment intensity in oncology scales with disease severity (Fig. 1). The Cox-guided gating mechanism provides an alternative to learned black-box routing by offering a transparent, statistically principled routing strategy: patient assignments are derived directly from familiar clinical features (e.g., age, cancer type, stage) rather than from opaque neural networks, enabling risk-aligned expert specialization. We position this approach as one principled point in the design space that prioritizes clinical transparency and risk-aligned specialization. Throughout this work, risk refers specifically to statistical clinical risk as quantified by the Cox proportional hazards model: a patient's risk score is the linear predictor (log hazard ratio) computed as a data-driven weighted combination of baseline clinical covariates, where higher scores correspond to greater instantaneous hazard of the event (death). The Cox model learns these weights directly from training data, providing a statistically principled, data-driven risk quantification that is used for patient stratification and expert routing.

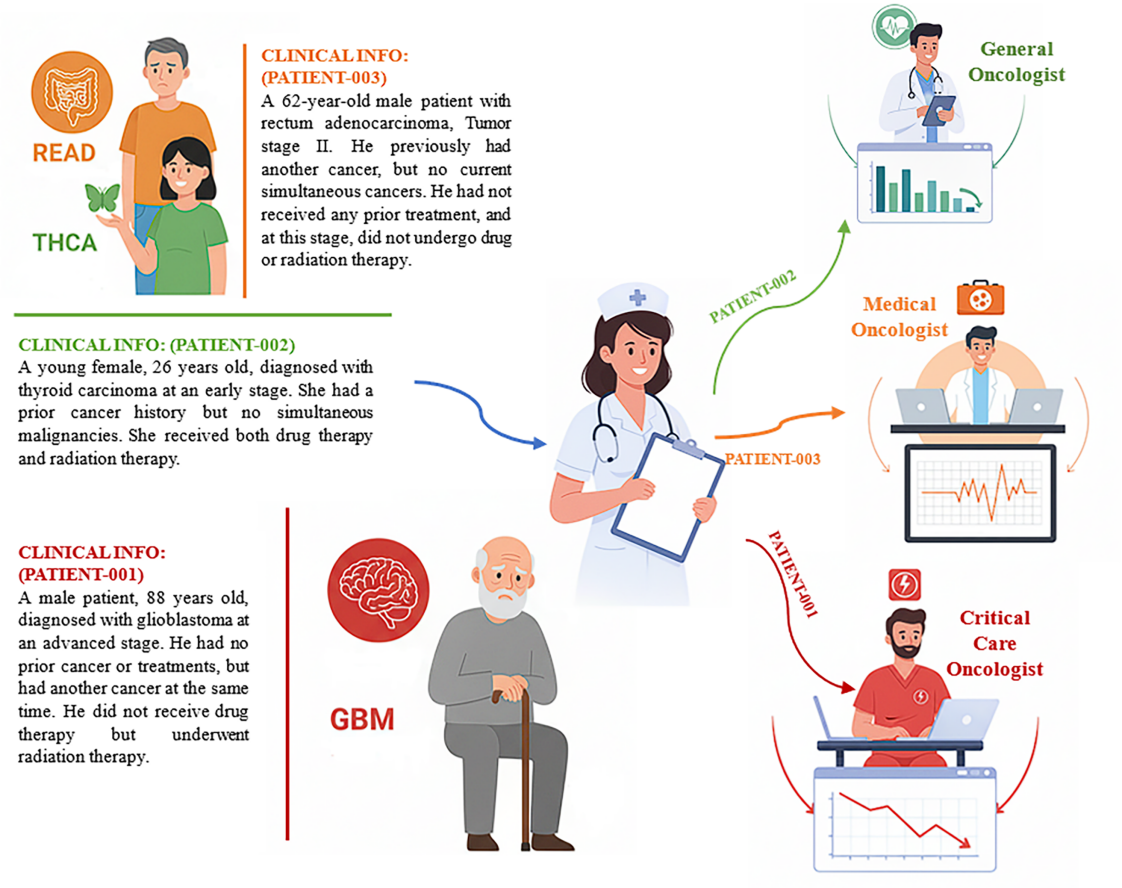

Figure 1: Conceptual overview of CoxGuided-SE. The clinical scenario illustrates how patients of different risk severity are routed to appropriate specialist care, conceptually mirroring the CoxGuided-SE framework where Cox-derived risk scores guide patient-to-expert assignment.

Within each risk stratum, CoxGuided-SE trains a dedicated deep expert network for survival prediction. We implement these experts using the same architecture as MultiSurv's [9] unimodal model for fair comparison. Notably, most recent deep learning approaches to pan-cancer survival have emphasized multimodal integration. MultiSurv [9] combined clinical, multi-omics, and histopathology features, while Fan et al. [10] employed attention-based fusion of clinical and molecular data. Subsequent work has explored tensor-based fusion strategies [11], histology-genomic integration [12,22], and diverse architectures including transformer-based [29], meta-learning [30], and interpretable multimodal models [31]. Studies focusing on individual modalities have predominantly examined mRNA expression [23,32–35], while other molecular data types, particularly miRNA and CNV, remain underexplored as standalone predictors, despite evidence linking CNV-driven genomic instability to downstream transcriptional dysregulation and tumor progression [36,37]. This exploration leaves open the question of how different data modalities individually benefit from architectural innovations such as expert specialization. In this work, we systematically evaluate each modality independently to isolate the effects of Cox-guided gating, providing insights into where expert specialization yields the greatest advantage. This design allows us to systematically test a core hypothesis: the benefit of expert specialization increases with data dimensionality and heterogeneity. We evaluate CoxGuided-SE across 33 cancer types using four modalities independently (clinical, mRNA, miRNA, CNV), demonstrating that expert specialization yields the greatest benefit for high-dimensional genomic data while maintaining computational efficiency.

The key contributions of this work can be listed as follows:

- **Interpretable Gating:** We introduce a novel Cox-based gating mechanism for mixture-of-experts survival models that derives data-driven risk strata from a CPH model. This avoids reliance on learned or heuristic routing rules; patients are assigned to experts via a transparent, statistically clinical risk score. By routing patients according to data-driven Cox-derived risk scores computed from clinical prognostic variables, experts receive more homogeneous subpopulations, simplifying the learning task, particularly for high-dimensional molecular data.
- **Risk-Stratified Expert Assignment:** To our knowledge, CoxGuided-SE is the first MoE survival model that routes patients based on clinical risk severity (low, intermediate, aggressive).
- **Balanced Expert Utilization:** The quantile-based stratification scheme guarantees balanced expert utilization by construction for any number of experts (K = 2, 3, 4, 5), preventing degenerate cases (e.g., one expert receiving nearly all patients) without requiring auxiliary losses or manual tuning. It provides a simple, effective solution to the expert collapse problem commonly encountered in hard gating MoEs.
- **Clinical Cox-Derived Modality-Independent Gating:** The Cox-guided routing mechanism operates identically regardless of the downstream prediction modality. While expert architectures are tailored to each data type (clinical, mRNA, miRNA, CNV), the patient stratification process remains unchanged, always derived from clinical covariates via the same Cox model.

Our extensive experiments on TCGA pan-cancer data show that this approach yields competitive or superior performance relative to strong baselines accuracy on challenging high-dimensional modalities (mRNA, miRNA, CNV), while maintaining excellent calibration and efficiency. By evaluating performance on individual molecular modalities (mRNA, miRNA, CNV) and clinical features separately, we demonstrate that high-dimensional, heterogeneous omics data benefit most profoundly from expert specialization, while simpler architectures adequately model low-dimensional clinical data. We believe this work takes an important step toward practical, trustworthy AI for precision oncology.

2 Related Work

Survival prediction has progressed through several methodological generations: from classical Cox proportional hazards models [6], to machine learning approaches such as random survival forests [38], and more recently to deep learning architectures like DeepSurv [39], MTLR [8], and multimodal frameworks [9,10]. Each stage has expanded the modeling capacity for complex, non-linear interactions in high-dimensional clinical and molecular data. Most recently, mixture-of-experts (MoE) frameworks have emerged as a promising solution to patient heterogeneity, offering adaptive routing of individuals to specialized subnetworks.

Recent frameworks have applied MoE concepts to oncology, yet they differ fundamentally in experimental scope, expert definition, and routing logic. Gene-MOE [23] utilizes a sparse, top-k gated architecture pre-trained on 33 cancer types, demonstrating the utility of self-supervised pre-training for learning generalizable RNA-seq features. However, survival evaluation is restricted to 14 cancer types, the framework is designed exclusively for mRNA expression, and no interpretability analysis of routing decisions or expert specialization was reported. MoME [20] and SurMoE [21] are designed as cancer-specific models, validated on 3 and 5 individual cancer cohorts, respectively. Their strength lies in capturing within-cancer heterogeneity through modality-biased expert design, MoME progressively reduces modality gaps through iterative cross-modal encoding, while SurMoE links expert behaviour to tissue patch clusters and gene pathway groups. However, neither learns a unified pan-cancer representation, and both derive routing signals from fused intermediate features rather than interpretable clinical variables. UMPSNet [22] introduces clinical metadata into routing through a CLIP-based text encoder, encouraging cancer-type specialization across 5 jointly trained cancer types. While this represents a step toward clinically informed routing, the gating mechanism remains tied to learned text projections rather than transparent clinical variable contributions, and the framework requires 10 experts for only 5 cancer types. Most recently, Morrill et al. [24] demonstrated that expert expressivity, not routing alone, drives calibration improvements in MoE survival models, achieving strong results on public clinical benchmarks. Their Personalized-MoE uses 8 experts with learned soft routing and load-balancing regularization. However, expert specialization is restricted to the output prediction head (all experts share a common backbone), and empirical validation is limited to two tabular clinical datasets without pan-cancer or molecular data evaluation. Notably, the definition of an "expert" varies widely across these studies: MoME and SurMoE treat experts as modality-biased architectural variants, UMPSNet aligns experts with cancer types, while Personalized-MoE defines experts as distribution generators. Across all reviewed frameworks, explicit stratification based on clinical prognostic severity remains uncommon.

A common limitation across many of these methods is that the gating mechanisms operate as black boxes. In Gene-MOE, SurMoE, and MoME, routing is determined by learned neural networks that discover latent patterns from data, making the decision of "which expert handles this patient" opaque. Even UMPSNet, which uses text prompts for gating, relies on learned projections to generate mixing weights. Direct integration of data-driven clinical risk stratification paradigms into MoE expert routing remains relatively underexplored in current survival modeling literature. Consequently, many existing MoE models partition data based on hidden feature statistics or fusion needs, rather than clinically interpretable prognostic severity. Clinical adoption may be limited when AI prognostic tools do not provide understandable reasoning behind predictions, despite the growing use of deep learning in clinical oncology applications [40].

Collectively, these observations highlight several open directions in the current MoE survival literature. To our knowledge, few frameworks have demonstrated unified training across all 33 TCGA cancer types with a small number of experts. In addition, most routing mechanisms derive their signal from the same feature space used for prediction (e.g., molecular inputs or fused representations), rather than from a separate

clinical information stream. Finally, while load-balancing regularization is widely used to encourage expert utilization, alternative designs that promote balanced specialization through the routing mechanism itself remain underexplored.

Cox-guided routing is a clinical-covariate-based choice that prioritizes auditability and stable expert utilization. However, learned routing can be advantageous when (i) clinical covariates are weak, missing, or noisy, (ii) heterogeneity is better explained by molecular or imaging features than baseline severity, or (iii) the risk structure is strongly non-linear or multi-factorial and not well captured by a linear Cox model. Multimodal gating can additionally exploit cross-modal interactions (e.g., when clinical stage modulates the prognostic meaning of a genomic alteration). We therefore view CoxGuided-SE as one point in a broader MoE design space that trades some flexibility for transparency, reproducibility, and clinician-aligned routing.

3 Method

3.1 Model Overview

Unlike traditional supervised survival models that train a single network across all patients, CoxGuided-SE decomposes the prediction task by partitioning the patient space into risk-homogeneous subgroups via a frozen Cox model, then training independent expert networks on each subgroup. This design separates the routing decision (derived from clinical features) from the survival prediction function (operating on molecular features or clinical feature), creating independent expert networks that each learn from distinct patient subgroups.

The proposed CoxGuided-SE is a hard gated MoE survival model that integrates Cox-based risk stratification with expert subnetworks adapted from the MultiSurv architecture. Unlike soft gated MoE approaches that blend outputs from multiple experts, CoxGuided-SE employs deterministic hard gating, routing each patient to a single specialized expert based on their clinical risk profile. This design emphasizes interpretability and computational efficiency while maintaining the core MoE principle of specialized expert networks. The model consists of two main components: an interpretable clinical gating mechanism and K expert subnetworks trained on risk-stratified patient subsets.

The proposed model comprises two phases as illustrated in (Fig. 2):

- *Phase I (Clinical Cox Model):* A clinical-only Cox proportional-hazards model is fitted on the training set to compute patient-specific risk scores and determine quantile-based risk thresholds. The Cox model is then frozen (parameters fixed) and used as the gating mechanism throughout Phase II.
- *Phase II (Expert Training):* Using the frozen Cox model from Phase I, each patient is deterministically assigned to a specialized expert network based on their risk stratum (e.g., low, intermediate, high). All K experts are trained simultaneously on their respective risk-stratified patient subsets, and each patient's final survival prediction is obtained exclusively from their assigned expert.

3.2 Cox-Guided Clinical Gating

CoxGuided-SE employs a deterministic, clinically interpretable gating mechanism that assigns each patient to one of K expert subnetworks based solely on a Cox proportional-hazards model [6] fitted to baseline clinical variables. This gating module operates independently of molecular features and therefore provides transparency and reproducibility. The Cox model was chosen for gating because it produces a single interpretable risk score as a transparent linear combination of clinical covariates, enabling clinically auditable patient routing. Non-linear alternatives would improve flexibility but sacrifice the transparency that is central to the framework's design.

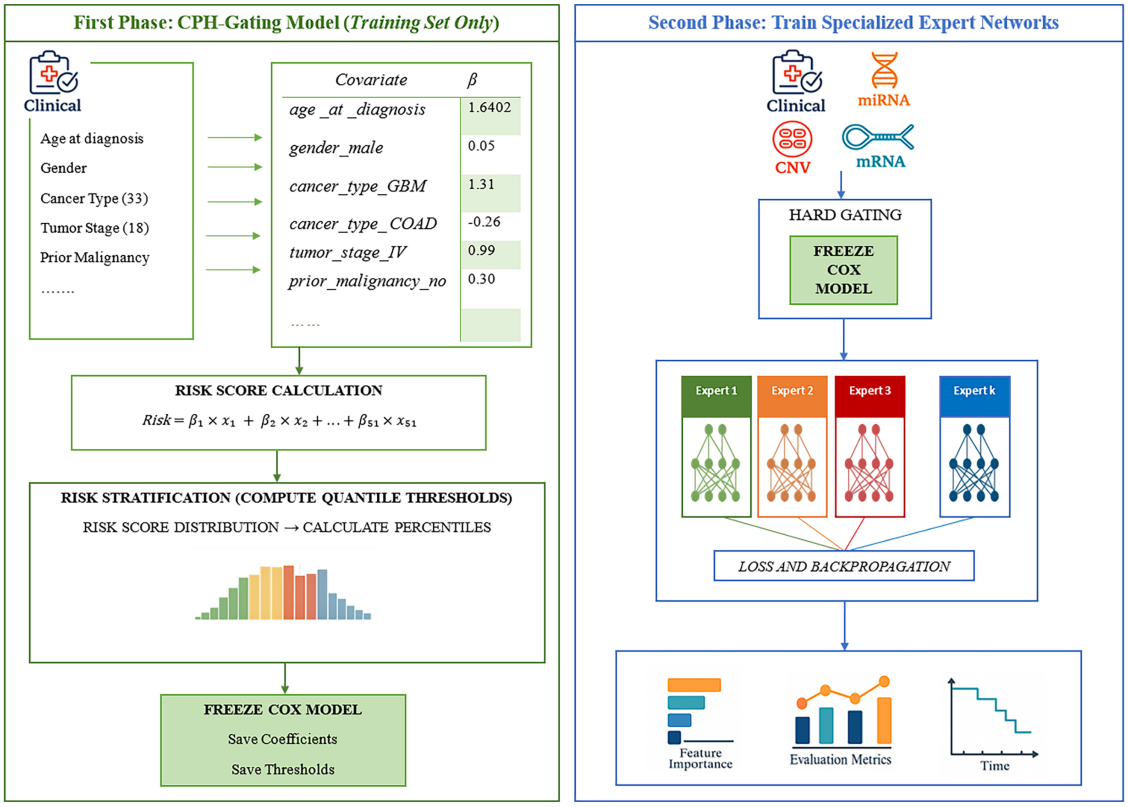

Figure 2: Schematic of the proposed CoxGuided-SE architecture. Phase I (left): a Cox proportional hazards model is fitted on clinical covariates to compute patient risk scores (Eqs. (1)–(4)) and quantile-based stratification thresholds (Eq. (5)). Phase II (right): the frozen Cox model routes each patient to a specialized expert network via deterministic hard gating (Eqs. (6) and (7)), and each expert is trained exclusively on its assigned risk stratum (Eq. (8)).

For patient $i \in \{1, \ldots, N\}$, let $x_i^{(\text{clin})} \in \mathbb{R}^p$ denote the vector of p clinical covariates, with observed follow-up time $T_i \in \mathbb{R}^+$ and event indicator $\delta_i \in \{0, 1\}$ (where 1 indicates an observed event and 0 indicates censoring). The Cox model estimates the log-risk for patient i as:

$$r_i = \beta^\top x_i^{(\text{clin})}, \tag{1}$$

where $r_i \in \mathbb{R}$ is the log-risk score and $\beta \in \mathbb{R}^p$ are the regression coefficients. This log-risk is associated with the hazard function:

$$h(t \mid x_i^{(\text{clin})}) = h_0(t) \exp(r_i) \tag{2}$$

where $h_0(t)$ is the baseline hazard function. The coefficients β are estimated by minimizing the negative Cox partial log-likelihood. For uncensored events ($\delta_i = 1$), the partial log-likelihood is:

$$\mathcal{L}_{\text{Cox}}(\beta) = -\sum_{i:\delta_i=1} \left[r_i - \log \left(\sum_{\ell \in \mathcal{R}_i} \exp(r_\ell) \right) \right] \tag{3}$$

where $\mathcal{R}_i = \{\ell : T_\ell \geq T_i\}$ is the risk set at time T_i (the set of patients still at risk when patient i experiences an event). To handle multicollinearity arising from one-hot encoded categorical clinical features across 33 cancer types, the model is fitted with $\ell 2$ ridge regularization ($\lambda = 0.1$), selected as the minimum penalization

required for stable convergence and reliable coefficient estimation. The resulting risk score provides a unified, standardized patient ordering across all 33 cancer types: the same Cox model with the same coefficients is applied to every patient regardless of cancer type, ensuring consistent and reproducible risk quantification.

We note that all clinical covariates used for gating are baseline variables recorded at diagnosis; no time-varying covariates are included. While the proportional hazards assumption may not hold perfectly for all covariates, the Cox model serves solely as a routing mechanism, and the expert subnetworks produce non-proportional survival predictions that are not constrained by this assumption.

A patient's clinical risk score is then computed as the weighted sum of their clinical features:

$$\text{RiskScore}_i = \sum_{j=1}^{p} \beta_j x_{ij}^{(\text{clin})} \in \mathbb{R} \tag{4}$$

where the sum is over all p clinical features for patient i, and $x_{ij}^{(\text{clin})}$ denotes the j-th component of $\mathrm{x}_i^{(\text{clin})}$. Variables associated with worse prognosis (e.g., older age, advanced stage, aggressive histology) contribute positively to the risk score, whereas features associated with more indolent disease (e.g., low-risk cancer types) contribute negatively.

3.3 Quantile-Based Risk Stratification

Quantile-based stratification addresses a fundamental challenge in hard gating MoE architectures: expert collapse. In learned routing mechanisms, the gating network often converges to activate only a subset of experts, leaving others undertrained [41]. This self-reinforcing pattern, where favored experts receive more training signals and become increasingly preferred, typically requires auxiliary load-balancing losses to mitigate.

CoxGuided-SE avoids this issue by deterministically partitioning patients into fixed quantiles of the training risk-score distribution. For a total of K experts ($\mathrm{K} \in \mathbb{N}$), each expert receives approximately 1/K of the training cohort, ensuring balanced expert utilization by design. Fig. 3 illustrates this balanced allocation for K = 2, 3, and 4 experts. This design prevents collapse without auxiliary losses while preserving the interpretability benefits of deterministic hard gating.

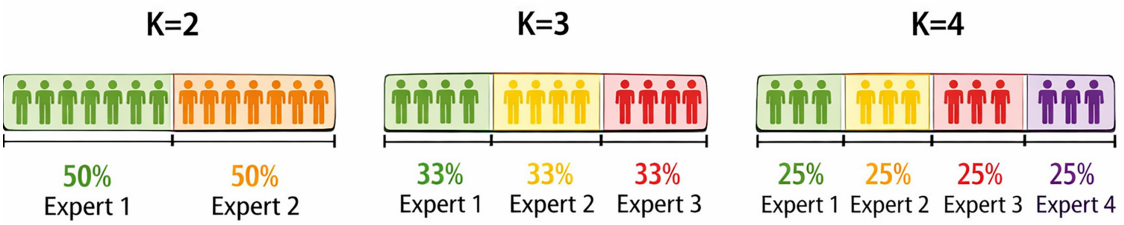

Figure 3: Quantile-based stratification ensures balanced expert utilization across varying numbers of experts.

Let $F\colon \mathbb{R} \to [0,1]$ denote the empirical cumulative distribution function (CDF) of the training-set risk scores, and $F^{-1}\colon (0,1) \to \mathbb{R}$ the corresponding quantile (inverse CDF) function. For each $k \in \{1, \ldots, K-1\}$, we define the quantile thresholds as:

$$q_k = F^{-1}\left(\frac{k}{K}\right) \in \mathbb{R} \tag{5}$$

with boundary conditions $q_0 = -\infty, q_K = +\infty$. These $K-1$ thresholds divide the risk-score space into K mutually exclusive and exhaustive strata, from lowest to highest risk.

Each patient $i \in \{1, \ldots, N\}$ is deterministically routed to exactly one expert based on their clinical risk score. Define the binary gating indicator $\pi_{ik} \in \{0,1\}$, where $\pi_{ik} = 1$ indicates that patient i is assigned to expert k, and 0 otherwise. Let $\Pi \in \{0,1\}^{N \times K}$ denote the binary assignment matrix such that each row corresponds to a single patient and each column to an expert. Hard gating requires that each patient be routed to exactly one expert, which is equivalent to enforcing a one-hot constraint:

$$\sum_{k=1}^{K} \pi_{ik} = 1, \forall i \in \{1, \ldots, N\} \tag{6}$$

Equivalently, each row of Π contains exactly one nonzero entry. We denote the i-th row as $\boldsymbol{\pi}_i^\top = [\pi_{i1}, \ldots, \pi_{iK}]$. The expert assignment is governed by the following rule:

$$\pi_{ik} = \begin{cases} 1, & \text{if } q_{k-1} \leq \text{RiskScore}_i < q_k \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

This enforces a one-hot gating vector $\boldsymbol{\pi}_i \in \{0,1\}^K$, where exactly one element is 1 and all others are 0. For instance, if $K = 3$, the thresholds q_1 and q_2 represent the 33rd and 67th percentiles of the training risk scores. Patients are assigned as follows:

- If $\text{RiskScore}_i < q_1 \Rightarrow \pi_{i1} = 1$ [Expert 1: Low Risk]
- If $q_1 \leq \text{RiskScore}_i < q_2 \Rightarrow \pi_{i2} = 1$ [Expert 2: Intermediate Risk]
- If $\text{RiskScore}_i \geq q_2 \Rightarrow \pi_{i3} = 1$ [Expert 3: Aggressive Risk]

This assignment is used consistently during training, validation, and testing.

3.4 Expert Subnetworks and Survival Prediction

Each expert network receives raw modality-specific features as input: clinical data (9 categorical variables processed through embedding layers + 1 continuous variable), mRNA (1000-dimensional continuous vector), miRNA (1881-dimensional continuous vector), or CNV (2000-dimensional categorical vector processed through embedding layers). Each expert produces a 30-dimensional output vector, where each element represents the conditional probability of surviving a one-year interval given survival to the beginning of that interval. The cumulative product of these conditional probabilities yields the patient's predicted survival curve spanning 30 years (details in Supplementary Material).

Each expert subnetwork in CoxGuided-SE adopts the unimodal architecture of MultiSurv [9], a deep survival model designed for discrete time-to-event pan-cancer prediction. MultiSurv was chosen as the expert backbone because it offers stable optimization behavior and strong empirical performance across TCGA modalities. Using MultiSurv as the expert network ensures that improvements observed in CoxGuided-SE arise from the gating mechanism rather than architectural differences. All expert subnetworks follow the original MultiSurv design, including fully connected layers with batch normalization, ReLU activation, and discrete-time survival outputs [9]. Full discrete-time survival and loss function details, including conditional survival probabilities, per-patient negative log-likelihood formulation, and gradient isolation under hard gating, are provided in the Supplementary Material.

The key distinction from standard MultiSurv [9] training lies in the hard gating mechanism. Due to hard gating, each patient's loss backpropagates only through their assigned expert. Patient i contributes loss exclusively to the expert k for which $\pi_{ik} = 1$. If $\pi_{ik} = 0$ for a given expert, that expert receives no gradient signal from patient i. The total loss is:

$$\mathcal{L}_{\text{total}} = \sum_{k=1}^{K} \sum_{i:\pi_{ik}=1} \ell_i^{(k)}, \tag{8}$$

where $\ell_i^{(k)}$ denotes the discrete-time negative log-likelihood survival loss for patient i evaluated through expert k (full derivation in Supplementary Material), ensuring that each expert is trained exclusively on its designated risk-stratified subset. This creates K independent optimization pathways, producing specialized models that focus on distinct regions of the prognostic risk space.

3.5 Model Interpretability and Expert Specialization

This gating strategy provides: (i) Deterministic routing: identical clinical inputs always map to the same expert, (ii) Clinically auditable logic: assignments depend solely on risk scores derived from transparent Cox coefficients, (iii) Expert specialization: each subnetwork receives a more homogeneous subset of patients, and (iv) Feature-level interpretability: because routing is fixed, each expert's molecular feature importance (see Feature Importance and Expert Specialization Analysis) can be unambiguously attributed. Unlike SoftMax gating, which obscures routing decisions behind a learned neural function, Cox-guided hard gating connects deep learning with data-driven clinical risk quantification, enabling reproducible and interpretable expert specialization.

Expert-specific feature importance was quantified using permutation importance. For each feature, values were randomly shuffled and the resulting C-td drop was measured. To ensure stability and reduce variance from random permutation order, each feature was permuted 100 times and the importance scores were averaged across repeats. This repeated-permutation approach provides variance-robust estimates without requiring computationally expensive bootstrap procedures, following established guidelines for permutation importance [42]. To quantify non-redundancy of learned representations, Jaccard similarity was computed between top-N feature sets for each expert pair:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{9}$$

Analysis was conducted for $N = 100$ showing divergent feature priorities across experts. We note that the interpretability provided by CoxGuided-SE supports model transparency and hypothesis generation rather than causal inference or direct clinical decision-making. Feature importance identifies molecular features associated with predictive performance within each risk stratum, but does not establish causal relationships. Prospective clinical validation would be required before deployment in clinical decision support. A complete illustration of this interpretability pipeline, from clinical features to Cox risk score, expert assignment, survival prediction, and patient-level feature importance, is provided for representative patients in Supplementary Fig. S1.

4 Experimental Setup

All models were trained under identical experimental conditions. Detailed implementation settings and computational environment specifications are provided in the Supplementary Material.

4.1 Dataset and Preprocessing

This study used The Cancer Genome Atlas (TCGA) pan-cancer dataset [15] comprising 11,081 patients across 33 cancer types, with four data modalities: clinical (10 variables), mRNA expression (1000

features), miRNA expression (1881 features), and CNV (2000 features). Data partitioning (80/10/10 training/validation/testing stratified by cancer type), feature normalization, and preprocessing followed the MultiSurv study [9] to ensure comparability. Full preprocessing details for each modality are provided in the Supplementary Material.

For clinical features, categorical variables were processed differently by the two model components: in the Cox gating module, multi-level categorical features were one-hot encoded with reference-category dropping to satisfy Cox regression requirements, while in the expert subnetworks, categorical variables were integer-encoded and processed through learned embedding layers, following the MultiSurv architecture. The continuous feature (age at diagnosis) was min–max scaled to $[0, 1]$.

Survival labels (time-to-event and censoring status) were obtained from the TCGA clinical data through the GDC Data Portal, following standardized definitions of overall survival used across prior pan-cancer studies [9,10].

4.2 Comparative Models

To contextualize CoxGuided-SE against established survival models, we compare it against a range of representative models spanning traditional statistical baselines and recent state-of-the-art deep learning approaches. CPH [6]: A classical linear survival model serving as the foundation for our gating strategy. MTLR [8]: A widely used deep survival method with strong performance in non-linear risk modeling. MultiSurv [9]: A modern discrete-time deep survival framework and one of the most competitive pan-cancer predictors; its unimodal subnetworks are also used as the backbone for expert architectures in CoxGuided-SE. Fan et al. (2023) [10]: A recent deep learning model evaluated across the same TCGA modalities used in this study, providing an up-to-date comparison for unimodal performance. To the best of our knowledge, MultiSurv and Fan et al. (2023) are the only existing studies that report results across all four unimodal settings (clinical, mRNA, miRNA, and CNV) used in our work. This makes them particularly valuable for direct benchmarking and highlights the novelty of our systematic evaluation across underexplored modalities. These baselines collectively cover classical, non-linear, modern deep learning, and recent pan-cancer survival methods, ensuring a comprehensive and fair performance evaluation.

A number of MoE-based models have recently been proposed for survival analysis, but most are unsuitable for the unimodal benchmarking focus of our study. Most existing MoE-based survival models are explicitly designed for multimodal integration. MoME [20], SurMoE [21], and UMPSNet [22] are inherently multimodal architectures that fuse information across data types; they cannot be directly applied to unimodal evaluation and do not report performance on individual modalities separately. By contrast, Gene-MOE [23], while unimodal, relies on extensive self-supervised pre-training on genomic data before survival fine-tuning. Applying Gene-MOE to our setting would require pre-training four separate foundation models (one per modality), introduce substantial computational overhead and confound the comparison by mixing pre-training benefits with architectural effects. Furthermore, Gene-MOE was evaluated for survival prediction on only 14 of the 33 cancer types. Among recent MoE survival architectures, Personalized-MoE [24] is the most suitable model: it is a general-purpose MoE framework applicable to any input modality without requiring pre-training, employs state-of-the-art learned soft routing with load-balancing regularization, and represents the current standard for MoE-based survival prediction. Together, CoxGuided-SE and Personalized-MoE represent two distinct points in the MoE design space: deterministic Cox-derived routing vs. learned soft routing with load-balancing regularization. Comparing these approaches on the same pan-cancer benchmark provides insight into whether transparent, risk-aligned routing can achieve competitive performance relative to flexible learned routing, an open question in MoE survival modelling.

4.3 Evaluation Metrics

Model performance was assessed using two complementary metrics, consistent with the evaluation protocol of MultiSurv [9] and computed using the pycox library [43]. Discrimination was measured using Antolini's adjusted time-dependent concordance index (C-td) [44], which quantifies the model's ability to correctly rank patients by predicted risk: higher predicted risk should correspond to earlier observed events. Unlike Harrell's C-index, Antolini's C-td permits patient risk rankings to vary across time points, making it appropriate for non-proportional discrete-time survival models such as CoxGuided-SE. Calibration was assessed using the Integrated Brier Score (IBS) [45], a proper scoring rule that summarizes the average squared distance between predicted survival probabilities and observed survival status over follow-up time (lower is better). The IBS was computed over 100 equidistant time points between the minimum and maximum event times in the test set, with the last quartile of time points excluded to avoid instability at late follow-up. Right-censored observations were handled via Kaplan-Meier-estimated inverse-probability-of-censoring weights (IPCW). The TCGA dataset contains one sample per patient (identified by submitter ID); no de-duplication was required. All metrics include 95% confidence intervals derived from 1000 bootstrap replicates using the percentile method [46].

5 Results

5.1 Determining the Optimal Number of Experts

To evaluate the effect of expert specialization, the number of experts (K) was varied from 2 to 5 across all data modalities (clinical, mRNA, miRNA, and CNV), with optimal K determined by balancing predictive performance, statistical stability, and parameter efficiency as shown in (Fig. 4a,b).

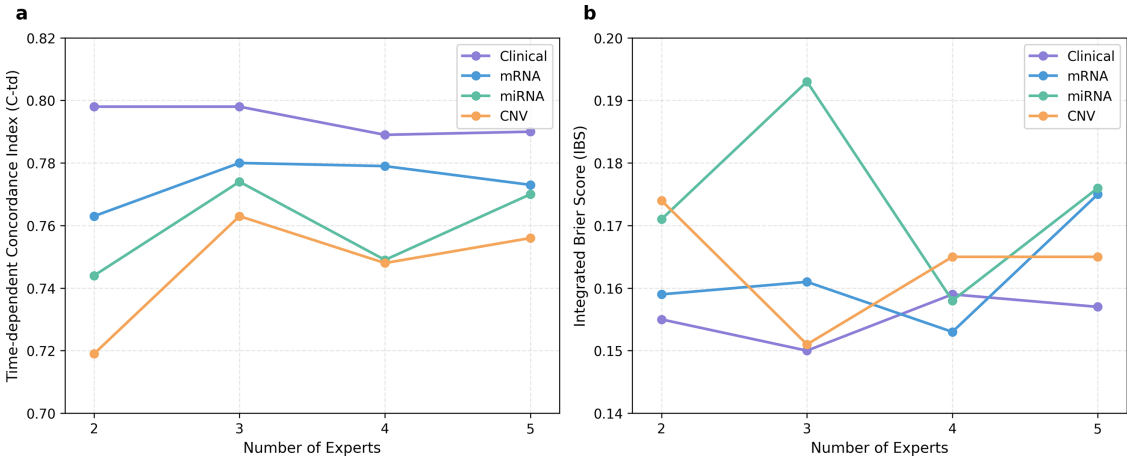

Figure 4: Effect of the number of experts on model performance. (**a**) Time-dependent concordance index (C-td) and (**b**) integrated brier score (IBS) for K = 2–5 across all modalities.

Increasing K from 2 to 3 yielded consistent improvements in both discrimination and calibration across modalities, indicating that moderate specialization enables the Cox-guided experts to capture latent heterogeneity in survival risk. Clinical data achieved peak discrimination at K = 3 (C-td = 0.798 [0.779–0.815]; IBS = 0.150 [0.138–0.165]). CNV showed substantial improvement, with C-td increasing from 0.719 (K = 2) to 0.763 (K = 3). miRNA achieved best discrimination at K = 3 despite modest calibration trade-offs (Supplementary Table S2). Across all modalities, K = 3 provided the largest gains relative to K = 2, after which performance stabilized. Beyond K = 3, marginal performance gains did not justify the increased model

complexity. For mRNA, C-td changed from 0.780 (K = 3) to 0.779 (K = 4), with overlapping confidence intervals (0.762–0.798 vs. 0.761–0.798) indicating no statistically significant difference. This modest change required adding one expert, increasing mRNA parameters from 14.1M to 18.8M (+33%). Similar proportional increases occurred across other modalities as shown in the Supplementary Table S3. Based on these results, K = 3 was adopted for all subsequent analyses, balancing accuracy, computational efficiency, and alignment with the conventional three-tier prognostic grouping (low, intermediate, high risk) used in clinical oncology.

5.2 Empirical Validation of Cox Risk Stratification

The Cox proportional hazards gating model was fitted on clinical covariates to derive interpretable patient-level risk strata. The model demonstrated robust discrimination (concordance index = 0.776) with stable convergence (Akaike Information Criterion = 44,302.0), confirming the presence of a reliable clinical risk signal in the training cohort. Quantile-based stratification produced balanced cohort partitioning across K strata; for K = 3, approximately one-third of training patients were allocated to each group (Low = 33.7%, Intermediate = 33.2%, Aggressive = 33.1%), ensuring adequate sample sizes for each expert while maintaining prognostic separation.

Regression coefficients (β) quantify each covariate's contribution to the log hazard ratio. Age at diagnosis yielded the largest positive coefficient (β = 1.63), corresponding to a hazard ratio of approximately 5.1 per standard deviation increase. Glioblastoma (GBM, β = 1.30, HR ≈ 3.7), Stage IV disease (β = 0.96, HR ≈ 2.6), and pancreatic adenocarcinoma (PAAD, β = 0.96, HR ≈ 2.6) also contributed substantial positive effects. Conversely, indolent malignancies showed strong negative coefficients: prostate adenocarcinoma (PRAD, β = −1.06, HR ≈ 0.35), thyroid carcinoma (THCA, β = −0.88, HR ≈ 0.41), and thymoma (THYM, β = −0.87, HR ≈ 0.42) (Fig. 5). These covariate effects align with established oncologic prognostic factors, where age and advanced stage are consistently associated with increased mortality across cancer types, while thyroid and prostate carcinomas are recognized for favorable long-term outcomes [19].

To illustrate the routing mechanism, consider a 75-year-old Black male patient with stage IV lung adenocarcinoma (LUAD), prior systemic treatment, a previous malignancy, and baseline radiation therapy. The risk score is computed as the sum of Cox coefficients multiplied by feature values (Fig. 5). Risk-increasing contributions include: normalized age (1.63 × 0.75 = 1.22), stage IV disease (0.96), LUAD cancer type (0.29), prior treatment (0.17), Black race (0.12), and male gender (0.05). The previous malignancy (−0.28) and radiation therapy (−0.15) provide modest risk reduction. The cumulative risk score of 2.38 exceeds the Aggressive-risk threshold (≥0.9), routing this patient to the Aggressive-risk expert.

Risk score thresholds were established at 0.2 and 0.9, corresponding to the 33rd and 67th percentiles of the training-set linear predictor distribution. These thresholds were applied unchanged to the held-out test set to define three risk strata (Fig. 6a). Kaplan–Meier analysis of test-set patients demonstrated clear, monotonic survival separation across Low-, Intermediate-, and Aggressive-risk groups over 20 years of follow-up (Fig. 6b). The log-rank test confirmed highly significant survival differences among the three strata (χ^2 = 296.88, df = 2, $p < 0.001$), indicating that the observed survival differences are highly unlikely to occur by chance. Event rates increased progressively across strata: 8.7% in Low-risk, 32.0% in Intermediate-risk, and 57.5% in Aggressive-risk groups (n = 1092 test patients), validating the clinical relevance of the Cox-based stratification. Nineteen test-set patients with zero recorded follow-up time were retained in the analysis to avoid selection bias; these patients are included in group totals but are consumed at t = 0 in the Kaplan–Meier estimation.

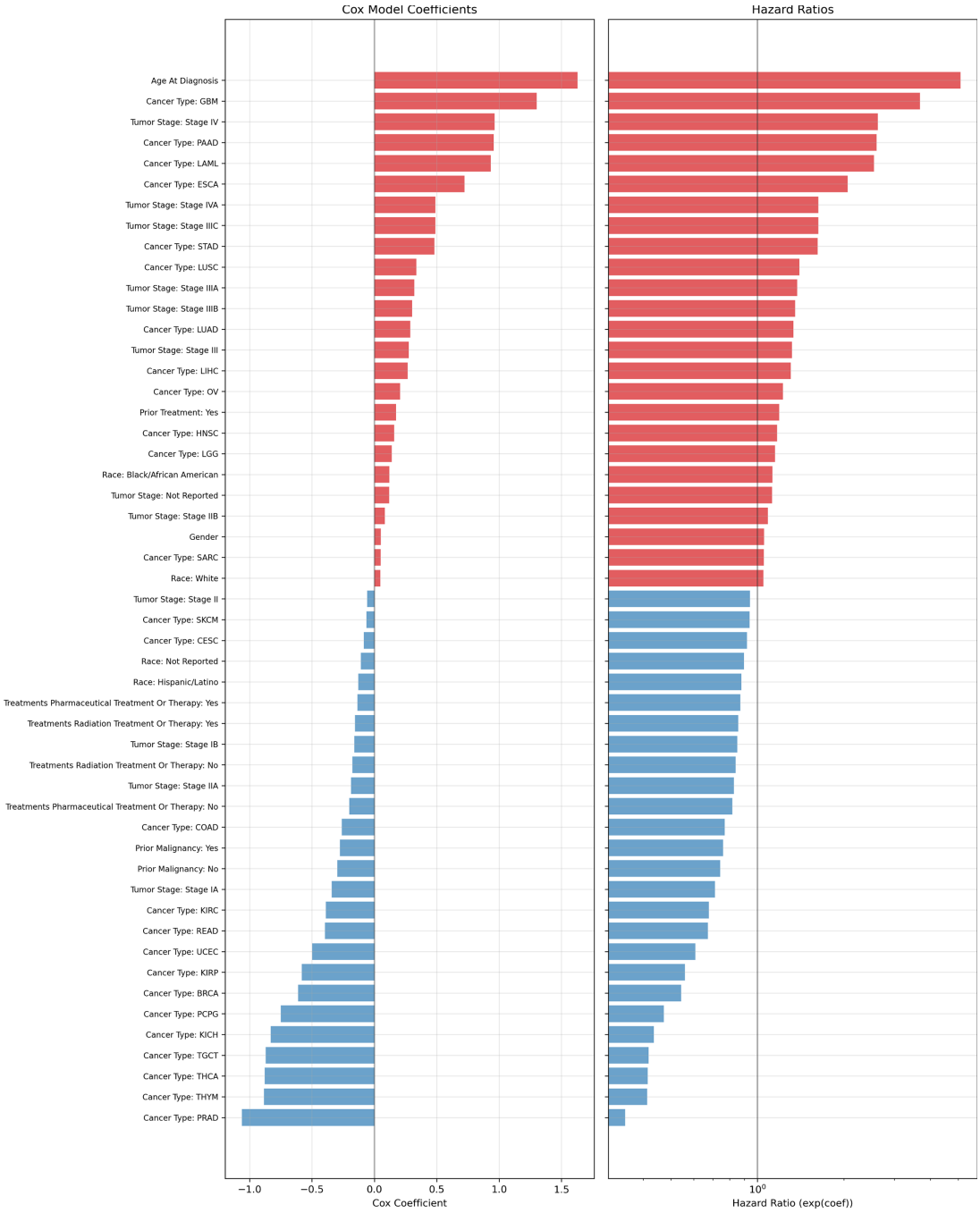

Figure 5: Empirical validation of Cox-based risk stratification; Cox regression coefficients and hazard ratios for clinical covariates.

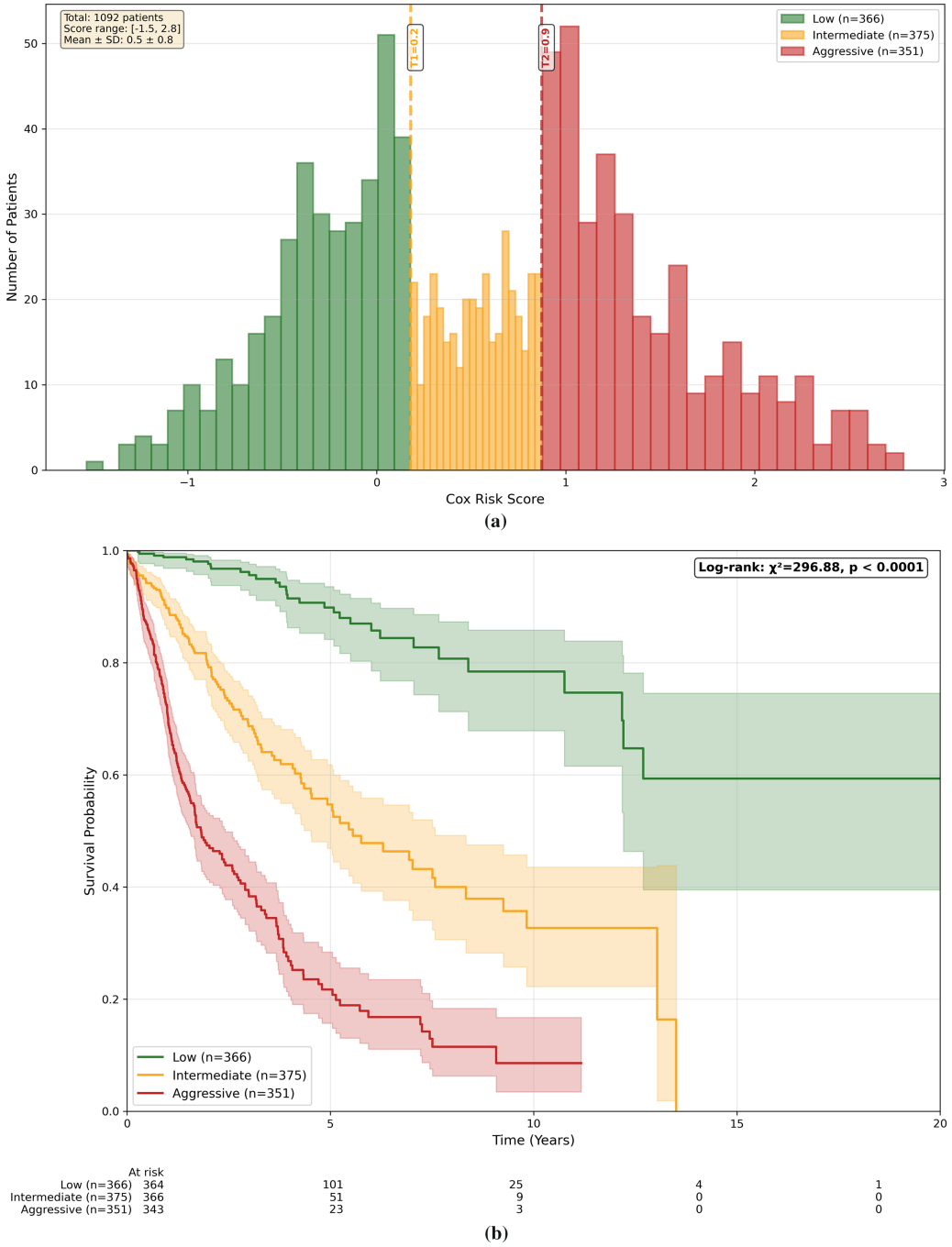

Figure 6: Empirical validation of Cox-based risk stratification; (**a**) risk score distribution with quantile thresholds (K = 3). (**b**) Kaplan–Meier curves demonstrating significant survival separation with risk table.

Sensitivity Analysis of Cox Model Configuration

To evaluate the robustness of Cox gating, we conducted ablation experiments examining two key design choices: clinical covariate selection and threshold configuration.

Covariate Ablation: Table 1 shows the effect of reducing clinical covariates from the full 10-variable model to minimal subsets. Using age at diagnosis alone, the single strongest prognostic factor [47], achieved reasonable discrimination (C-td = 0.770 for clinical, 0.696 for mRNA, 0.691 for miRNA, 0.627 for CNV), demonstrating that even minimal clinical information provides useful stratification. Adding cancer type improved performance across all modalities, with the largest gains for CNV. The full covariate model consistently achieved the highest discrimination, with CNV showing the greatest sensitivity to covariate richness $\Delta = 0.136$ from age-at-diagnosis-only to the full clinical model (all covariates). These results suggest that high-dimensional molecular features benefit most from informative clinical stratification, while the framework remains functional even with limited clinical data.

Table 1: Effect of Cox model covariate subsets on discrimination (C-td). Bold indicates best performance; 95% confidence intervals in parentheses.

Model	Age-Only	Age + Cancer Type	Age + Cancer Type + Tumor Stage	All Clinical Covariates
Clinical	0.770 (0.752–0.789)	0.773 (0.755–0.792)	0.767 (0.748–0.787)	**0.798 (0.779–0.815)**
mRNA	0.696 (0.673–0.718)	0.736 (0.715–0.754)	0.764 (0.747–0.781)	**0.78 (0.762–0.798)**
miRNA	0.691 (0.669–0.713)	0.725 (0.703–0.748)	0.757 (0.738–0.776)	**0.774 (0.757–0.793)**
CNV	0.627 (0.593–0.658)	0.721 (0.694–0.750)	0.757 (0.732–0.779)	**0.763 (0.746–0.782)**

Threshold Ablation: Holding the Cox covariate set fixed to the full clinical model. Table 2 examines sensitivity to quantile threshold selection for K = 3 experts. Alternative configurations (25th/75th, 40th/60th, 15th/85th percentiles) yielded C-td values within approximately 0.02–0.05 of the selected 33rd/67th thresholds across all modalities, indicating that performance is robust to reasonable threshold choices. The 33rd/67th configuration achieved the highest or near-highest discrimination for all modalities. Beyond discrimination, threshold selection directly affects expert utilization balance, which is central to preventing expert collapse, a failure mode in hard-gating MoE architectures where a subset of experts receives insufficient training data, leading to degenerate routing and underutilization [48]. The 15th/85th configuration concentrates 70% of patients in the Intermediate-risk expert, effectively reducing the model to near-single-expert behaviour for most patients. The 40th/60th configuration allocates only 20% to the Intermediate-risk expert, leaving it undertrained relative to the other two. Only the 33rd/67th configuration guarantees equal utilization (33%/33%/33%), ensuring that every expert receives sufficient training signal for robust specialization. These results support the adopted thresholds as balancing discrimination performance and stable expert utilization without requiring auxiliary load-balancing losses.

Table 2: Effect of quantile threshold selection on discrimination (C-td) and expert utilization. Bold indicates best performance; 95% confidence intervals in parentheses.

Model	25th/75th	40th/60th	15th/85th	33rd/67th
Clinical	0.763 (0.743–0.781)	0.783 (0.764–0.801)	0.758 (0.736–0.777)	**0.798 (0.779–0.815)**
mRNA	0.761 (0.743–0.781)	0.750 (0.727–0.770)	0.761 (0.743–0.781)	**0.78 (0.762–0.798)**
miRNA	0.745 (0.725–0.764)	0.773 (0.755–0.793)	0.739 (0.718–0.759)	**0.774 (0.757–0.793)**
CNV	0.748 (0.722–0.775)	0.751 (0.726–0.776)	0.717 (0.689–0.745)	**0.763 (0.746–0.782)**
Expert Utilization	25%/50%/25%	40%/20%/40%	15%/70%/15%	**33%/33%/33%**

5.3 Performance across Modalities

To evaluate prognostic performance, the proposed CoxGuided-SE model (K = 3) was benchmarked against several models. All models were trained using identical data partitions and hyperparameter configurations. Results are summarized in Table 3.

Table 3: Performance comparison of CoxGuided-SE against baseline and state-of-the-art methods across all data modalities. Bold indicates best performance; 95% confidence intervals in parentheses.

Metrics	Data	Models					
		CPH [6]	MTLR [8]	Fan et al. (2023) [10]	MultiSurv [9]	Personalized-MoE [24]	CoxGuided-SE
Ctd	Clinical	0.796 (0.779–0.813)	**0.811 (0.795–0.828)**	0.778 (0.751–0.802)	0.797 (0.78–0.816)	0.753 (0.733–0.774)	0.798 (0.779–0.815)
	mRNA	0.706 (0.680–0.729)	0.733 (0.709–0.755)	0.725 (0.694–0.757)	0.739 (0.716–0.762)	0.717 (0.695–0.741)	**0.78 (0.762–0.798)**
	miRNA	0.674 (0.649–0.699)	0.626 (0.599–0.652)	0.687 (0.655–0.720)	0.672 (0.646–0.699)	0.67 (0.647–0.695)	**0.774 (0.757–0.793)**
	CNV	0.570 (0.544–0.597)	0.589 (0.562–0.615)	0.613 (0.579–0.647)	0.537 (0.511–0.563)	0.563 (0.536–0.59)	**0.763 (0.746–0.782)**
IBS	Clinical	0.143 (0.135–0.154)	**0.136 (0.125–0.15)**	0.155 (0.141–0.170)	0.163 (0.149–0.18)	0.16 (0.152–0.173)	0.15 (0.138–0.165)
	mRNA	0.196 (0.178–0.21)	0.187 (0.168–0.204)	0.183 (0.166–0.203)	0.182 (0.165–0.199)	0.192 (0.173–0.207)	**0.161 (0.144–0.179)**
	miRNA	0.184 (0.170–0.199)	0.197 (0.18–0.215)	**0.180 (0.156–0.206)**	0.198 (0.177–0.213)	0.19 (0.174–0.205)	0.193 (0.169–0.212)
	CNV	0.214 (0.207–0.224)	0.221 (0.206–0.239)	0.217 (0.202–0.235)	0.216 (0.203–0.23)	0.223 (0.217–0.231)	**0.151 (0.139–0.166)**

For clinical data, CoxGuided-SE achieved C-td = 0.798, comparable to other deep learning methods. MTLR achieved the numerically highest discrimination (0.811), though differences among methods were minimal. This confirms that low-dimensional clinical covariates (10 variables) provide limited opportunity for specialized architectures to demonstrate an advantage, as expected. For mRNA expression, CoxGuided-SE achieved the highest discrimination (C-td = 0.780) and best calibration (IBS = 0.161), outperforming all baselines including MultiSurv (0.739) and MTLR (0.733). Partial separation of bootstrap intervals indicates reliable improvement. Substantial gains were observed for miRNA. CoxGuided-SE reached C-td = 0.774, substantially exceeding all baselines (next best: Fan et al. at 0.687). The non-overlapping confidence intervals indicate a statistically robust advantage.

The largest improvement occurred for CNV data. CoxGuided-SE achieved C-td = 0.763, substantially exceeding all other models (next best: Fan et al. at 0.613). IBS improved to 0.151, representing a 30% reduction compared to the best state-of-the-art. Confidence intervals were entirely separated, indicating clear statistical significance.

Compared to Personalized-MoE [24], a recent MoE model employing learned soft routing with load-balancing regularization, CoxGuided-SE demonstrated superior discrimination across all modalities: CNV (+35.5%), miRNA (+15.5%), mRNA (+8.8%), and clinical (+6.0%). Calibration was also improved for three of four modalities. These results suggest that clinically derived deterministic gating can outperform learned soft routing on heterogeneous pan-cancer data. Notably, CoxGuided-SE achieves these gains without requiring auxiliary load-balancing losses.

To formally verify the statistical significance of the reported performance differences, Friedman tests [49] were conducted across methods for each modality-metric combination, with all tests yielding $p < 0.001$. Post-hoc pairwise Wilcoxon signed-rank tests with Holm correction were applied to confirm the significance of CoxGuided-SE's improvement over the strongest competing method in each molecular modality [50,51]. For mRNA, CoxGuided-SE significantly outperformed MultiSurv in both C-td ($p < 0.001$) and IBS ($p < 0.001$). For miRNA, CoxGuided-SE significantly exceeded Fan et al. in C-td ($p < 0.001$). For miRNA calibration (IBS), CoxGuided-SE did not improve over Fan et al. (0.193 vs. 0.180), reflecting the discrimination–calibration trade-off observed under hard gating for this modality. For CNV, CoxGuided-SE significantly outperformed Fan et al. in both C-td ($p < 0.001$) and IBS ($p < 0.001$). For clinical data,

CoxGuided-SE achieved comparable performance to the strongest baselines, with overlapping confidence intervals across multiple methods (Table 3), consistent with the expectation that low-dimensional clinical features offer limited room for expert specialization.

To further contextualize performance, we compared CoxGuided-SE against MultiSurv trained for 75 epochs (MultiSurv-75), representing the original study's full training protocol. Results are shown in Table 4.

Table 4: Performance comparison between MultiSurv-75 (full training) and CoxGuided-SE (15 epochs). Bold indicates best performance; 95% confidence intervals in parentheses.

Metric	Model	Clinical	mRNA	miRNA	CNV
Ctd	**MultiSurv-75**	**0.801 (0.784–0.819)**	0.754 (0.731–0.776)	0.717 (0.691–0.74)	0.592 (0.565–0.62)
	CoxGuided-SE	0.798 (0.779–0.815)	**0.78 (0.762–0.798)**	**0.774 (0.757–0.793)**	**0.763 (0.746–0.782)**
IBS	**MultiSurv-75**	0.152 (0.14–0.167)	0.184 (0.164–0.198)	0.2 (0.173–0.216)	0.212 (0.202–0.224)
	CoxGuided-SE	**0.15 (0.138–0.165)**	**0.161 (0.144–0.179)**	**0.193 (0.169–0.212)**	**0.151 (0.139–0.166)**

For clinical data, CoxGuided-SE showed comparable performance to MultiSurv-75, with overlapping confidence intervals (C-td: 0.798 vs. 0.801; IBS: 0.150 vs. 0.152). However, for high-dimensional modalities, CoxGuided-SE achieved superior performance despite using only 15 epochs compared to 75. For mRNA, CoxGuided-SE outperformed MultiSurv-75 (C-td: 0.780 vs. 0.754; IBS: 0.161 vs. 0.184). For miRNA, the improvement was more pronounced (C-td: 0.774 vs. 0.717), with non-overlapping confidence intervals. The largest gains occurred for CNV: CoxGuided-SE achieved C-td = 0.763 compared to 0.592 for MultiSurv-75, representing a +28.9% improvement. IBS improved from 0.212 to 0.151 (−28.8%). These results demonstrate that performance gains stem from architectural innovation (Cox-guided expert specialization) rather than extended training, as CoxGuided-SE achieves superior results with 5× fewer epochs.

To further assess prediction reliability beyond discrimination, calibration was evaluated by comparing predicted survival probabilities against Kaplan–Meier observed estimates at 1-, 3-, and 5-year horizons for CNV data (Fig. 7). At the 1-year horizon, both models showed reasonable calibration in the high-probability range where most predictions concentrate, though CoxGuided-SE tracked the diagonal more consistently. The advantage became more pronounced at longer horizons: at 3 and 5 years, CoxGuided-SE maintained closer agreement between predicted and observed survival across the full probability range, whereas MultiSurv-75 exhibited systematic deviation, particularly overestimating survival for higher-risk patients (predicted probabilities below 0.6). This pattern is consistent with the substantial IBS improvement reported in Table 4 (0.212 → 0.151) and confirms that CoxGuided-SE's gains extend beyond patient ranking to clinically meaningful probability estimation.

From a clinical perspective, well-calibrated survival probabilities are essential for informed treatment planning and patient communication. For example, a predicted 5-year survival probability of 0.6 should correspond to approximately 60% of patients in that group surviving to 5 years. Fig. 7 shows that CoxGuided-SE achieves this correspondence more reliably than MultiSurv-75, particularly at 3- and 5-year horizons where treatment decisions and prognostic counselling are most relevant. In contrast, MultiSurv-75 exhibits greater deviation from perfect calibration, especially in the mid-probability range, which may lead to less reliable prognostic communication in clinical settings.

To mitigate the risk of unreliable estimates from small cohorts, cancer-specific evaluation was restricted to cancer types with at least 20 test samples. We note that CoxGuided-SE trains a unified pan-cancer model rather than cancer-specific models; each expert receives patients from multiple cancer types within its risk stratum, reducing the risk of overfitting to any single small cohort.

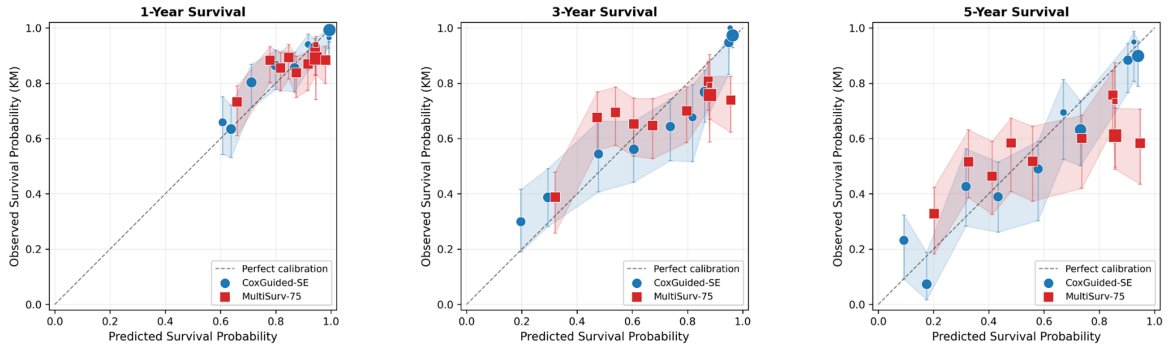

Figure 7: Calibration plots for CNV data comparing CoxGuided-SE and MultiSurv-75 at 1-, 3-, and 5-year survival horizons. Patients were grouped into deciles of predicted survival probability; observed survival was estimated using the Kaplan–Meier method within each bin. Points closer to the dashed diagonal indicate better calibration. Error bars and shaded regions represent 95% confidence intervals.

Fig. 8 presents cancer-specific performance for the CNV modality. CoxGuided-SE achieved higher C-td and lower IBS than MultiSurv-75 in the majority of cancer types (15 of 19 for C-td, 14 of 19 for IBS), reflecting effective expert specialization for high-dimensional molecular features. Notable improvements include: Kidney renal papillary cell carcinoma (KIRP) C-td: 0.723 → 0.938, Colon adenocarcinoma (COAD) 0.587 → 0.820, Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) 0.587 → 0.804, Breast invasive carcinoma (BRCA) 0.515 → 0.779, and PRAD 0.667 → 0.788. IBS improvements follow the same trend, indicating gains in both discrimination and calibration.

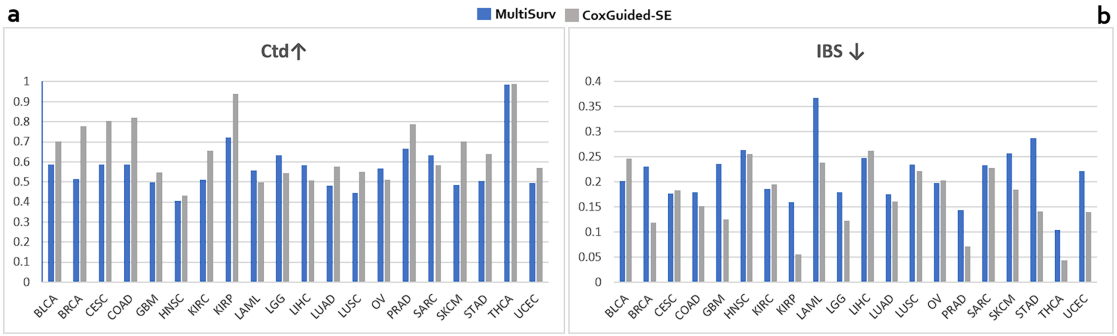

Figure 8: Cancer-specific survival prediction performance for CNV data. (**a**) C-td (higher is better). (**b**) IBS (lower is better) for CoxGuided-SE vs. MultiSurv-75 across TCGA cancer types with ≥20 samples.

However, CoxGuided-SE showed reduced CNV discrimination for a subset of cancers, including LGG (0.633 → 0.545), LIHC (0.583 → 0.508), and OV (0.565 → 0.512). Notably, LGG and LIHC showed a consistent pattern in which CoxGuided-SE underperformed MultiSurv-75 in C-td across molecular modalities (mRNA, miRNA, CNV), despite achieving improved clinical prediction (Supplementary Tables S4–S7). A more nuanced pattern was observed for LGG. Although discrimination decreased, calibration improved substantially across molecular modalities (mRNA IBS: 0.236 → 0.127; miRNA: 0.175 → 0.098; CNV: 0.179 → 0.123), indicating that CoxGuided-SE produced more accurate survival probability estimates even when patient ranking was less precise. In contrast, LIHC represented a clearer negative case, with reduced C-td and inconsistent calibration changes across modalities.

These findings are consistent with established disease biology. For example, LGG prognosis differs substantially across molecular subtypes defined by 1p/19q co-deletion status, which define biologically and

clinically distinct disease entities [52,53]. Similarly, survival heterogeneity in LIHC has been associated with distinct etiologic and molecular subtypes (e.g., viral vs. metabolic disease), which exhibit different genomic and clinical risk profiles [54,55]. In such settings, clinical risk–based partitions may group molecularly dissimilar patients within the same expert, thereby introducing confounding variance that undermines, rather than enhances, expert specialization.

In addition, within a unified pan-cancer model spanning 33 malignancies, cancer-specific prognostic signals may be diluted relative to features with broader cross-cancer relevance, and may therefore not appear prominently among top-ranked expert features. These results suggest that extending CoxGuided-SE with cancer-aware or hierarchical routing, where pan-cancer risk stratification is complemented by cancer-specific refinement, may improve performance in such settings. Together, these boundary cases clarify an important applicability condition: CoxGuided-SE is most effective when clinical severity provides a meaningful proxy for underlying molecular heterogeneity.

5.4 Expert Specialization and Feature Importance Analysis

To assess whether the Cox-guided gating mechanism encourages expert specialization rather than redundant learning, we quantified expert-specific feature usage using permutation importance and measured overlap across experts using pairwise Jaccard similarity. For each molecular modality, features were ranked by expert-specific permutation importance, and Jaccard indices were computed between expert pairs using the top-100 ranked features to evaluate robustness.

Across all molecular modalities, expert networks utilized divergent feature subsets, with consistently low Jaccard overlap (Table 5). These results indicate that experts assign different priorities to different features rather than converging on a common set of globally dominant features. Some degree of feature overlap across experts is expected, as certain molecular features carry prognostic value across risk strata; the key observation is that the relative importance rankings differ substantially across experts (Fig. 9). A permutation test (10,000 permutations) [56] further supports this interpretation: for several expert pairs, the observed overlap was significantly lower than expected under random feature assignment (e.g., miRNA Low ∩ Aggressive: $p = 0.028$; CNV Low ∩ Aggressive: $p = 0.006$).

Inspection of expert-specific feature rankings further illustrates this behavior (Fig. 9). Within each modality, different experts emphasize different subsets of high-ranked features, with minimal overlap among the top-100 ranked features across risk strata. This qualitative separation is consistent with the low Jaccard overlap quantified in Table 5. Together with the architectural guarantee that hard gating creates independent optimization pathways with no gradient sharing (Eq. (8)) and the substantial performance gains over single-model baselines, these results support genuine functional specialization across experts.

Table 5: Jaccard similarity indices for top-100 features across expert pairs. Values shown as Jaccard index (intersection/union). Low values indicate strong expert specialization.

Modality	Low ∩ Intermediate	Low ∩ Aggressive	Intermediate ∩ Aggressive	Average
mRNA	0.070 (13/187)	0.058 (11/189)	0.093 (17/183)	0.074
miRNA	0.031 (6/194)	0.005 (1/199)	0.026 (5/195)	0.021
CNV	0.015 (3/197)	0.020 (4/196)	0.026 (5/195)	0.020

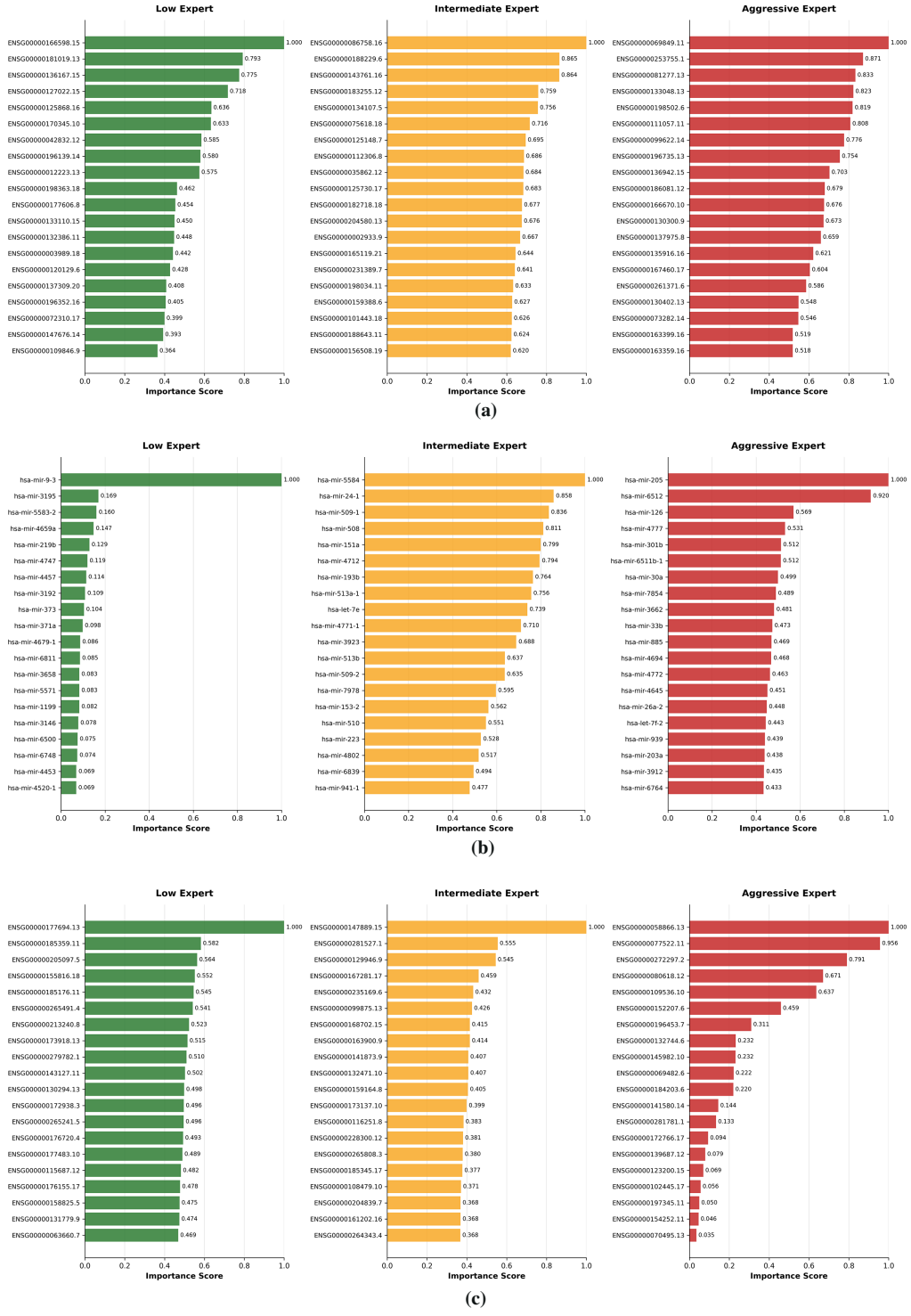

Figure 9: Expert-level feature specialization across modalities. The top-20 most important features identified for each of the three experts (low-, intermediate-, and aggressive-risk). (**a**) mRNA feature importance profiles. (**b**) miRNA feature importance profiles. (**c**) CNV feature importance profiles.

From a modeling perspective, this specialization provides a mechanistic explanation for the performance gains observed in high-dimensional modalities. By routing patients into clinically defined risk strata,

CoxGuided-SE reduces heterogeneity within each expert's training data, simplifying the learning problem and enabling each expert to model a narrower conditional distribution. Collectively, the experts cover a broader and more diverse region of the feature space than a single unified model, which helps explain the substantial improvements observed for high-dimensional molecular inputs.

To provide limited biological context for the observed expert specialization (Table 5), we examined selected representative top-ranked features from each expert. This inspection was qualitative rather than systematic. Some features identified by the model have previously been implicated in established cancer biology pathways reported in the literature. For example, ATP1B3, which ranked among the top features for the Aggressive-risk expert, has been reported to regulate gastric cancer progression via PI3K/AKT signaling [57]. Similarly, miR-205 is a canonical regulator of epithelial–mesenchymal transition (EMT) through targeting ZEB1 and ZEB2 [58], a key mechanism in cancer invasion and metastasis. This further aligns the Aggressive-risk expert's signature with a pro-invasive molecular phenotype. In the Intermediate-risk expert, CDKN2A maps to the CDK4/6–RB cell-cycle checkpoint axis, a pathway central to CDK4/6 inhibitor sensitivity mechanisms [59]. These observations are presented as illustrative examples only, and systematic biological validation of expert-specific signatures remains beyond the scope of this study and a direction for future work.

6 Discussion

This work introduced CoxGuided-SE, a time-to-event survival prediction framework that combines transparent Cox-based risk stratification with specialized expert networks. Across four data modalities, the proposed approach achieved notable improvements over established baselines and state-of-the-art, including CPH, MTLR, Fan et al. (2023), MultiSurv and Personalized-MoE, particularly for high-dimensional molecular data (mRNA, miRNA, CNV). Performance on clinical covariates remained comparable to existing methods, indicating that expert specialization provides the greatest benefit when patient heterogeneity and feature dimensionality are high. These findings are especially notable for CNV data, widely regarded as one of the most challenging modalities for survival modeling, where many deep learning approaches show limited or inconsistent gains. The CNV results suggest that expert specialization can effectively handle noisy, complex molecular profiles with non-linear associations to survival.

CoxGuided-SE instantiates a clinically interpretable routing principle, risk-severity decomposition, that is particularly well aligned with pan-cancer settings where baseline clinical variables provide a strong global prognostic signal. In regimes where this signal is weak or unavailable, or where heterogeneity is predominantly molecular, more flexible learned or multimodal gating may yield better partitions, at the cost of reduced auditability. These framing positions Cox-guided routing as one point in a broader design space that trades flexibility for transparency, reproducibility, and clinician-aligned stratification.

Our approach differs fundamentally from prior MoE survival models in both interpretability and scalability. While methods like MoME [20] and SurMoE [21] achieve strong performance, they rely on training separate models for each cancer type (e.g., one model for BRCA, another for LUAD). This cancer-specific paradigm does not leverage shared biological signals across malignancies. In contrast, CoxGuided-SE trains a single unified model across all 33 cancer types. By routing patients based on clinical risk severity rather than cancer labels, the model enables knowledge transfer across malignancies. This cross-cancer learning capability, combined with the efficiency of hard gating, explains why CoxGuided-SE achieves strong performance on high-dimensional molecular data (CNV, miRNA) with only 15 training epochs. The result is a more data-efficient and scalable solution.

Furthermore, CoxGuided-SE achieves competitive or superior performance with only three experts, compared to the 10 experts required by UMPSNet. This efficiency stems from the principled risk-based partitioning: rather than requiring many experts to capture arbitrary data patterns, our clinical gating naturally separates patients into prognostically meaningful subgroups that align with established oncology practice.

The comparison with Personalized-MoE [24] is particularly instructive, as CoxGuided-SE achieved superior discrimination across all modalities despite Personalized-MoE employing learned soft routing and load-balancing regularization (Section 5.3). This performance gap highlights a key insight: effective expert specialization does not necessarily require highly flexible routing mechanisms, but can instead be achieved through clinically grounded, deterministic stratification of the patient population.

CoxGuided-SE provides interpretability at two complementary levels. At the routing level, expert assignment is deterministic and reproducible, derived directly from a Cox-based clinical risk score constructed from standard clinical variables. This allows expert selection to be traced to interpretable, data-driven Cox coefficients applied to standard clinical variables rather than latent neural routing functions. At the feature-utilization level, permutation importance reveals expert-specific molecular drivers, with consistently low Jaccard overlap confirming that experts learn complementary, non-redundant representations. Together, these properties align with clinical decision-support requirements, where transparency, traceability, and reproducibility are essential for regulatory acceptance and clinical adoption. Importantly, these explanations are intended to support interpretability and hypothesis generation rather than direct therapeutic decision-making, which requires prospective validation.

The benefits of expert specialization scaled with input dimensionality and biological complexity. Clinical covariates, with limited variation across a small number of features, provided little room for specialization, consistent with the comparable performance between CoxGuided-SE and baselines on this modality. In contrast, high-dimensional omics data contain latent structure, subpopulations, and nonlinear effects that are effectively captured by the gating mechanism. This pattern demonstrates that expert specialization is particularly effective for noisy, heterogeneous molecular data that challenge conventional architectures.

Quantile-based stratification inherently prevents expert collapse by guaranteeing balanced expert utilization without auxiliary losses (Section 3.3). The hard gating design also offers computational benefits: unlike soft gating, which evaluates all experts per patient, deterministic routing requires inference through only a single expert, reducing computational cost proportionally to the number of experts. Empirically, CoxGuided-SE completed training in 12–25 min across modalities (15 epochs), compared to 49–60 min for MultiSurv (75 epochs), achieving 2.1–5.2× faster training while delivering superior discrimination and calibration (Supplementary Material). This efficiency is particularly relevant for deployment in resource-constrained clinical settings.

Although this study was designed to isolate the architectural effects of Cox-guided routing under controlled unimodal settings, we conducted limited feasibility experiments to evaluate whether the gating mechanism extends to multimodal inputs. Using a simple early-fusion strategy with the Cox gate unchanged, the clinical + mRNA configuration achieved a C-td of approximately 0.794 (0.777–0.811), while full multimodal integration achieved approximately 0.786 (0.767–0.804). These results are comparable to reported multimodal pan-cancer benchmarks while using substantially simpler architectures and shorter training schedules. Importantly, performance did not degrade substantially when additional modalities were introduced, suggesting that Cox-guided routing remains compatible with multimodal survival modelling. However, these experiments were not intended as a comprehensive multimodal optimization study; modality-aware fusion strategies and cross-modal interaction modelling remain important directions for future work.

This study has several limitations that suggest directions for future investigation. First, validation scope: The analysis was restricted to TCGA data; although the method was tested across four modalities and 33 cancer types, external validation on independent cohorts (e.g., CPTAC) is recommended to confirm data-level generalizability beyond the TCGA population. Second, modality integration: The present analysis evaluated each data modality independently to isolate expert specialization effects; extending the framework to fully optimized multimodal integration represents an important next step. Third, hard gating trade-offs: While the hard gating mechanism maximizes interpretability by assigning each patient to a single expert, it creates sharp decision boundaries where patients near risk thresholds could be routed differently based on minor clinical variations. Soft gating mechanisms, where predictions are weighted averages across experts, could improve stability for borderline cases at the cost of reduced interpretability. Additionally, incorporating uncertainty quantification into the expert predictions, for example, to flag patients near risk thresholds where routing decisions are most sensitive, represents a promising direction for enhancing the framework's reliability in clinical deployment. Fourth, linearity of gating: The Cox-based risk score relies on linear feature combinations, which may not capture complex non-linear interactions between clinical variables. Investigating hierarchical or non-linear gating functions (e.g., survival trees or neural networks) while preserving interpretability represents a promising direction. Fifth, stratification optimality: The quantile-based stratification creates equally-sized risk groups, which may not be optimal for all cancer types or data distributions. Adaptive or unsupervised clustering methods could identify more natural, variably-sized patient subgroups. Finally, while feature importance provides model-level interpretability, comprehensive biological validation, including systematic pathway enrichment analysis, cancer-type–specific investigation, and functional experiments, remains an important direction for future interdisciplinary research.

While this study focuses on oncology, the core idea, using an interpretable risk model to define expert routing, may be applicable to other time-to-event prediction settings where clinically meaningful stratification is standard.

7 Conclusion

This study addressed the challenge of pan-cancer survival prediction under pronounced patient heterogeneity by introducing CoxGuided-SE, a mixture-of-experts framework that embeds clinical risk stratification directly into model architecture. By redefining expert specialization as a clinically interpretable decomposition of the prognostic risk space rather than a black-box routing problem, the proposed approach enables deterministic, interpretable, and balanced expert assignment without reliance on learned gating networks or auxiliary regularization. Extensive evaluation on TCGA pan-cancer data demonstrated that this risk-guided specialization yields substantial and statistically robust improvements for high-dimensional molecular modalities, including mRNA, miRNA, and CNV, while offering limited but competitive benefits for low-dimensional clinical features.

These findings provide both a performant modeling framework and a principled explanation of when and why expert specialization is effective in survival analysis, advancing interpretable and trustworthy clinical artificial intelligence. By aligning deep learning architectures with principled survival modeling, CoxGuided-SE establishes a transparent foundation for risk-aware prognostic modeling and demonstrates methodological generalizability across data modalities, cancer types, and expert configurations without requiring task-specific modifications.

Acknowledgement: We would like to thank Ministry of Higher Education, Malaysia and Hadhramaut Establishment for Human Development for their funding and facility support to carry out this research successfully.

Funding Statement: This research was funded by the Ministry of Higher Education, Malaysia under the Fundamental Research Grant Scheme (FRGS/1/2019/ICT02/UKM/02/9).

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, designed and implemented the proposed model, performed experiments, analyzed the data, and wrote the first draft of the paper, Manal Mohammed AL-Tamimi; Supervised the study, Siti Norul Huda Sheikh Abdullah, Mohammad Khatim Hasan and Mohammed Azmi Al-Betar; Editing the manuscript, Siti Norul Huda Sheikh Abdullah, Mohammad Khatim Hasan, Mohammed Azmi Al-Betar, Maw Shin Sim and Abdulrahman Mohammed AL-Tamimi; Fund acquisition, Siti Norul Huda Sheikh Abdullah; Validation and investigation, Maw Shin Sim and Abdulrahman Mohammed AL-Tamimi. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The TCGA pan-cancer dataset used in this study is publicly available from the Genomic Data Commons (GDC) Data Portal (https://portal.gdc.cancer.gov/). The code for CoxGuided-SE is available from the authors upon request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Supplementary Materials: The supplementary material is available online at https://www.techscience.com/doi/10.32604/cmes.2026.079891/s1, Discrete-Time Survival and Loss Function. Implementation Details. Supplementary Table S1. Training Time Comparison; Table S2. Effect of Number of Experts (K) on Model Performance; Table S3. Model Parameters by Number of Experts; Table S4. Cancer-Specific Performance (Clinical Data); Table S5. Cancer-Specific Performance (mRNA Data); Table S6. Cancer-Specific Performance (miRNA Data); Table S7. Cancer-Specific Performance (CNV Data); Fig. S1. Representative Case Examples.

References

1. Collins GS, Chester-Jones M, Gerry S, Ma J, Matos J, Sehjal J, et al. Clinical prediction models using machine learning in oncology: challenges and recommendations. BMJ Oncol. 2025;4(1):e000914. doi:10.1136/bmjonc-2025-000914.
2. Raza SA. Emerging trends in global cancer epidemiology. Cancers. 2025;17(9):1483. doi:10.3390/cancers17091483.
3. Nygren P. Precision cancer medicine 2025: some concerns. Acta Oncol. 2025;64:1202–4. doi:10.2340/1651-226x.2025.44604.
4. Tran D, Nguyen H, Pham VD, Nguyen P, Nguyen Luu H, Minh Phan L, et al. A comprehensive review of cancer survival prediction using multi-omics integration and clinical variables. Brief Bioinform. 2025;26(2):bbaf150. doi:10.1093/bib/bbaf150.
5. Xu Q, Adam A, Abdullah A, Bariyah N. Advanced deep learning approaches in detection technologies for comprehensive breast cancer assessment based on WSIs: a systematic literature review. Diagnostics. 2025;15(9):1150. doi:10.3390/diagnostics15091150.
6. Cox DR. Regression models and life-tables. J R Stat Soc Ser B Methodol. 1972;34(2):187–202. doi:10.1111/j.2517-6161.1972.tb00899.x.
7. Yousefi S, Amrollahi F, Amgad M, Dong C, Lewis JE, Song C, et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. Sci Rep. 2017;7(1):11707. doi:10.1038/s41598-017-11817-6.
8. Fotso S. Deep neural networks for survival analysis based on a multi-task framework. arXiv:1801.05512. 2018.
9. Vale-Silva LA, Rohr K. Long-term cancer survival prediction using multimodal deep learning. Sci Rep. 2021;11(1):13505. doi:10.1038/s41598-021-92799-4.
10. Fan Z, Jiang Z, Liang H, Han C. Pancancer survival prediction using a deep learning architecture with multimodal representation and integration. Bioinform Adv. 2023;3(1):vbad006. doi:10.1093/bioadv/vbad006.
11. Chen J, Liu P, Chen C, Su Y, Zuo E, Li M, et al. TDMFS: tucker decomposition multimodal fusion model for pan-cancer survival prediction. Artif Intell Med. 2025;162(3):103099. doi:10.1016/j.artmed.2025.103099.

12. Hu Y, Li X, Yi Y, Huang Y, Wang G, Wang D. Deep learning-driven survival prediction in pan-cancer studies by integrating multimodal histology-genomic data. Brief Bioinform. 2025;26(2):bbaf121. doi:10.1093/bib/bbaf121.
13. Waqas A, Tripathi A, Ahmed S, Mukund A, Farooq H, Johnson JO, et al. Self-normalizing multi-omics neural network for pan-cancer prognostication. Int J Mol Sci. 2025;26(15):7358. doi:10.3390/ijms26157358.
14. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell. 2018;173(2):291–304.e6. doi:10.1016/j.cell.2018.03.022.
15. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The cancer genome atlas pan-cancer analysis project. Nat Genet. 2013;45(10):1113–20. doi:10.1038/ng.2764.
16. Mohd Yunos RI, Ab Mutalib NS, Khoo JS, Saidin S, Ishak M, Syafruddin SE, et al. Whole genome sequencing of Malaysian colorectal cancer patients reveals specific druggable somatic mutations. Front Mol Biosci. 2023;9:997747. doi:10.3389/fmolb.2022.997747.
17. Balan D, Kampan NC, Plebanski M, Abd Aziz NH. Unlocking ovarian cancer heterogeneity: advancing immunotherapy through single-cell transcriptomics. Front Oncol. 2024;14:1388663. doi:10.3389/fonc.2024.1388663.
18. Ab Mutalib NS, Othman SN, Mohamad Yusof A, Abdullah Suhaimi SN, Muhammad R, Jamal R. Integrated microRNA, gene expression and transcription factors signature in papillary thyroid cancer with lymph node metastasis. PeerJ. 2016;4(3):e2119. doi:10.7717/peerj.2119.
19. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell. 2018;173(2):400–16.e11. doi:10.1016/j.cell.2018.02.052.
20. Xiong C, Chen H, Zheng H, Wei D, Zheng Y, Sung JJY, et al. MoME: mixture of multimodal experts for cancer survival prediction. arXiv:2406.09696. 2024.
21. Zhang W, Xu W, Chen T, Sakal C, Li X. Integrating images and genomics for multi-modal cancer survival analysis via mixture of experts. Inf Fusion. 2026;126(13):103521. doi:10.1016/j.inffus.2025.103521.
22. Zhang B, Li S, Jian J, Ren X, Zhao Z, Guo L, et al. From single-cancer to pan-cancer prognosis. Am J Pathol. 2025;195(10):1869–84. doi:10.1016/j.ajpath.2025.06.006.
23. Meng X, Li X, Yang Q, Dai H, Qiao L, Ding H, et al. Gene-MOE a sparsely gated cancer diagnosis and prognosis framework exploiting pan-cancer genomic information. IEEE Trans Comput Biol Bioinform. 2025;22(2):514–27. doi:10.1109/TCBBIO.2024.3524209.
24. Morrill T, Puli A, Megjhani M, Park S, Zemel R. Let the experts speak: improving survival prediction & calibration via mixture-of-experts heads. arXiv:2511.09567. 2025.
25. Bertero L, Massa F, Metovic J, Zanetti R, Castellano I, Ricardi U, et al. Eighth edition of the UICC classification of malignant tumours: an overview of the changes in the pathological TNM classification criteria—what has changed and why? Virchows Arch. 2018;472(4):519–31. doi:10.1007/s00428-017-2276-y.
26. A predictive model for aggressive non-Hodgkin's lymphoma. [cited 2025 Jan 1]. Available from: https://www.nejm.org/doi/full/10.1056/NEJM199309303291402.
27. Amin MB, Greene FL, Edge SB, Compton CC, Gershenwald JE, Brookland RK, et al. The eighth edition AJCC cancer staging manual: continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. CA A Cancer J Clin. 2017;67(2):93–9. doi:10.3322/caac.21388.
28. Brierley J, O'Sullivan B, Asamura H, Byrd D, Huang SH, Lee A, et al. Global consultation on cancer staging: promoting consistent understanding and use. Nat Rev Clin Oncol. 2019;16(12):763–71. doi:10.1038/s41571-019-0253-x.
29. Li X, Min W, Chen J, Wu J, Wang S. TransVCOX: bridging transformer encoder and pre-trained VAE for robust cancer multi-omics survival analysis. In: Proceedings of the 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2023 Dec 5–8; Istanbul, Turkiye. p. 1254–9.
30. Cho HJ, Shu M, Bekiranov S, Zang C, Zhang A. Interpretable meta-learning of multi-omics data for survival analysis and pathway enrichment. Bioinformatics. 2023;39(4):btad113. doi:10.1093/bioinformatics/btad113.

31. Azher ZL, Vaickus LJ, Salas LA, Christensen BC, Levy JJ. Development of biologically interpretable multimodal deep learning model for cancer prognosis prediction. In: Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing; 2022 Apr 25–29; Virtual. p. 636–44.
32. Meng X, Wang X, Zhang X, Zhang C, Zhang Z, Zhang K, et al. A novel attention-mechanism based cox survival model by exploiting pan-cancer empirical genomic information. Cells. 2022;11(9):1421. doi:10.3390/cells11091421.
33. Cai H, Liao Y, Zhu L, Wang Z, Song J. Improving cancer survival prediction via graph convolutional neural network learning on protein-protein interaction networks. IEEE J Biomed Health Inform. 2024;28(2):1134–43. doi:10.1109/JBHI.2023.3332640.
34. Thedinga K, Herwig R. A gradient tree boosting and network propagation derived pan-cancer survival network of the tumor microenvironment. iScience. 2021;25(1):103617. doi:10.1016/j.isci.2021.103617.
35. Pavageau M, Rebaud L, Morel D, Christodoulidis S, Deutsch E, Massard C, et al. DeepOS: pan-cancer prognosis estimation from RNA-sequencing data. 2021 [cited 2025 Jan 1]. Available from: https://www.medrxiv.org/content/10.1101/2021.07.10.21260300v1.
36. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, et al. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013;45(10):1134–40. doi:10.1038/ng.2760.
37. Ali Hassan NZ, Mokhtar NM, Kok Sin T, Mohamed Rose I, Sagap I, Harun R, et al. Integrated analysis of copy number variation and genome-wide expression profiling in colorectal cancer tissues. PLoS One. 2014;9(4):e92553. doi:10.1371/journal.pone.0092553.
38. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat. 2008;2(3):841–60. doi:10.1214/08-aoas169.
39. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol. 2018;18(1):24. doi:10.1186/s12874-018-0482-1.
40. Gong W, Vaishnani DK, Jin XC, Zeng J, Chen W, Huang H, et al. Evaluation of an enhanced ResNet-18 classification model for rapid on-site diagnosis in respiratory cytology. BMC Cancer. 2025;25(1):10. doi:10.1186/s12885-024-13402-3.
41. Zhou Y, Lei T, Liu H, Du N, Huang Y, Zhao V, et al. Mixture-of-experts with expert choice routing. arXiv:2202.09368. 2022.
42. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics. 2010;26(10):1340–7. doi:10.1093/bioinformatics/btq134.
43. Kvamme H, Borgan Ø, Scheel I. Time-to-event prediction with neural networks and cox regression. arXiv:1907.00825. 2019.
44. Antolini L, Boracchi P, Biganzoli E. A time-dependent discrimination index for survival data. Stat Med. 2005;24(24):3927–44. doi:10.1002/sim.2427.
45. Gerds TA, Schumacher M. Consistent estimation of the expected brier score in general survival models with right-censored event times. Biom J. 2006;48(6):1029–40. doi:10.1002/bimj.200610301.
46. Rousselet GA, Pernet CR, Wilcox RR. The percentile bootstrap: a primer with step-by-step instructions in R. Adv Methods Pract Psychol Sci. 2021;4:1–10. doi:10.31234/osf.io/kxarf.
47. Kim SJ, Myong JP, Suh H, Lee KE, Youn YK. Optimal cutoff age for predicting mortality associated with differentiated thyroid cancer. PLoS One. 2015;10(6):e0130848. doi:10.1371/journal.pone.0130848.
48. Shazeer N, Mirhoseini A, Maziarz K, Davis A, Le Q, Hinton G, et al. Outrageously large neural networks: the sparsely-gated mixture-of-experts layer. arXiv:1701.06538. 2017.
49. Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Stat Assoc. 1937;32(200):675–701. doi:10.1080/01621459.1937.10503522.
50. Wilcoxon F. Individual comparisons by ranking methods. Biom Bull. 1945;1(6):80–3. doi:10.2307/3001968.
51. Holm S. A simple sequentially rejective multiple test procedure a simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6:65–70.
52. Cancer Genome Atlas Research Network. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. N Engl J Med. 2015;372(26):2481–98. doi:10.1056/nejmoa1402121.

53. Louis DN, Perry A, Reifenberger G, von Deimling A, Figarella-Branger D, Cavenee WK, et al. The 2016 world health organization classification of tumors of the central nervous system: a summary. Acta Neuropathol. 2016;131(6):803–20. doi:10.1007/s00401-016-1545-1.
54. Ally A, Balasundaram M, Carlsen R, Chuah E, Clarke A, Dhalla N, et al. Comprehensive and integrative genomic characterization of hepatocellular carcinoma. Cell. 2017;169(7):1327–41.e23. doi:10.1016/j.cell.2017.05.046.
55. Llovet JM, Kelley RK, Villanueva A, Singal AG, Pikarsky E, Roayaie S, et al. Hepatocellular carcinoma. Nat Rev Dis Primers. 2021;7(1):6. doi:10.1038/s41572-020-00240-3.
56. Marozzi M. Some remarks about the number of permutations one should consider to perform a permutation test. Statistica. 2004;64(1):193–201.
57. Li L, Feng R, Xu Q, Zhang F, Liu T, Cao J, et al. Expression of the β3 subunit of Na^+/K^+-ATPase is increased in gastric cancer and regulates gastric cancer cell progression and prognosis via the PI3/AKT pathway. Oncotarget. 2017;8(48):84285–99. doi:10.18632/oncotarget.20894.
58. Gregory PA, Bert AG, Paterson EL, Barry SC, Tsykin A, Farshid G, et al. The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. Nat Cell Biol. 2008;10(5):593–601. doi:10.1038/ncb1722.
59. O'Leary B, Finn RS, Turner NC. Treating cancer with selective CDK4/6 inhibitors. Nat Rev Clin Oncol. 2016;13(7):417–30. doi:10.1038/nrclinonc.2016.26.