 DOI: https://doi.org/10.12345/yzlcyxzz.v9i3.37308

A Systematic Review of Deep Representation Learning for Multi-omics Cancer Prognosis

Qitao Chen[1,2]

1. School of Public Health, Qilu Medical College, Shandong University, Jinan, Shandong, 250012, China
2. National Institute of Health and Medical Big Data, Shandong University, Jinan, Shandong, 25002, China

Abstract

The high incidence and mortality of malignant tumors pose persistent challenges to global healthcare systems. Traditional prognostic assessment relies on the TNM staging system, which inadequately reflects the molecular complexity of tumors. With the advancement of high-throughput sequencing technologies and artificial intelligence, deep learning-based survival prediction models integrating clinical and multi-omics data have emerged as a research hotspot. This review systematically summarizes the theoretical foundations, model architectures, and research progress in this field, with emphasis on the applications of deep feature learning in multimodal patient representation, individualized risk prediction, heterogeneity subtype identification, and disease evolution trajectory inference. Key challenges including data heterogeneity, sample size limitations, model interpretability, external validation, and ethical privacy concerns are thoroughly discussed. Future directions such as multi-center data standardization, explainable AI architectures, transfer learning, and federated learning are prospected, providing references for precision oncology and personalized clinical decision-making.

Keywords

Multi-omics data integration; Deep learning; Survival prediction; Precision medicine; Cancer prognosis

多组学癌症预后深度表征学习系统综述

陈齐涛 [1,2]

1. 山东大学齐鲁医学院公共卫生学院，中国 · 山东 济南 250012
2. 山东大学国家健康医疗大数据研究院，中国 · 山东 济南 250002

摘　要

恶性肿瘤的高发病率与高死亡率对全球医疗体系构成持续挑战。传统预后评估依赖TNM分期系统，难以充分反映肿瘤分子层面的复杂性。随着高通量测序技术与人工智能的发展，基于临床与多组学数据整合的深度学习生存预测模型成为研究热点。本文系统综述了该领域的理论基础、模型架构及研究进展，重点阐述了深度特征学习在患者多模态特征表征、个体化风险预测、异质性亚类识别及疾病演化轨迹推断中的应用。同时深入探讨了数据异质性、样本量限制、模型可解释性、外部验证及伦理隐私等关键挑战，并展望了多中心数据标准化、可解释AI架构、迁移学习与联邦学习等未来发展方向，为肿瘤精准医疗与个体化决策提供参考。

关键词

多组学数据整合；深度学习；生存预测；精准医疗；肿瘤预后

1 引言

恶性肿瘤是全球主要公共卫生问题之一，其高发病率与高死亡率持续对医疗体系构成挑战 [1]。根据全球癌症统，肺癌、乳腺癌与结直肠癌等仍为主要癌种负担来源。尽管手术、放疗、化疗、靶向治疗与免疫治疗等手段不断发展，但由于肿瘤高度的时空异质性及患者个体差异，临床疗效与生存结局仍存在显著差异。因此，构建准确可靠的生存预测模型，实现患者风险分层与个体化治疗决策支持，是当前临床亟需解决的核心问题。传统肿瘤预后评估主要依赖 TNM 分期系统与临床病理特征，这本质上是对疾病状态的“静态快照”，难以充分反映肿瘤分子层面的复杂性及疾病演化过程。随着高通量测序技术发展，体细胞突变、拷贝数变异、转录组、表观遗传等多组学数据广泛用于肿瘤机制研究与精准医疗。研究表明，驱动基因突变与关键通路改变与治疗反应、疾病进展及生存预后密切相关 [2-3]。

2 基于临床与多组学数据整合的生存预测模型理论基础及研究框架

2.1 生存分析基础

生存分析（Survival analysis）用于研究从起点事件到终点事件所经历的时间。经典方法包括 Kaplan-Meier 生存曲线、对数秩检验（Log-rank test）以及 Cox 比例风险模型（Cox proportional hazards model, Cox-PH）等。

Cox 模型假设协变量对风险函数的影响为线性，其定

义为：

$$h(t|X)=h_0(t)\exp(\beta^T x)$$

其中 $h(t|X)$ 为给定协变量 X 的风险函数，$h_0(t)$ 为基线风险函数，β 为回归系数向量。

然而，Cox 模型在处理高维组学数据时易面临维度灾难、共线性与非线性关系刻画不足等问题。为克服这些局限，研究者逐步引入深度学习方法以提升对复杂数据结构的建模能力。

2.2 深度学习生存预测模型

深度学习生存模型通过神经网络学习高层次特征表示，可处理高维、多模态数据并捕捉复杂非线性关系。当前主流架构正从简单的多层感知机（Multi-layer perceptron, MLP）向基于注意力机制（Attention mechanism）的 Transformer 架构演进，以捕捉长程依赖和模态间交互。表 1 总结了常用于生存预测的深度学习模型的架构及其特点。

Katzman 等提出 DeepSurv，通过多层感知机替代 Cox 模型的线性项，以 Cox 部分似然为损失函数，实现个体化风险预测。在 METABRIC 数据集上，DeepSurv 报告的一致性指数（Concordance index, C-index）为 0.654。Lee 等提出 DeepHit，将生存时间离散化并直接预测生存概率分布。此外，Kvamme 等系统讨论了神经网络与 Cox 回归结合的时间到事件（Time-to-event）预测框架。

表 1 常用于生存预测的深度学习模型

模型	架构特点	优点	缺点	适用性
DeepSurv	基于多层感知机的 Cox 模型扩展	捕捉非线性关系；保留 Cox 可解释性	仍依赖比例风险假设	适用于中等规模队列
DeepHit	离散时间生存模型，可处理竞争风险	无需比例风险假设；适合竞争风险	需离散时间；训练复杂	适用于复杂生存分析
Transformer-based	自注意力建模长程依赖	强交互建模；可解释性较好	参数量大；需大样本	适用于大规模多组学
GNN-based	图结构建模患者关系	利用患者相似性关系	图构建影响性能	适用于相似性网络预后
VAE-based	生成式潜变量表征学习	处理缺失；可量化不确定性	训练不稳定	适用于稀疏高维表征

2.3 多组学数据

临床数据通常为低维结构化特征，具有可解释性强、获取成本低等特点；而组学数据往往高维、稀疏、噪声大，且存在批次效应与缺失问题（表 2）。

表 2 临床数据和基因突变数据特点总结

数据类型	数据特点	生物学意义	适用性
临床特征	低维结构化（年龄、性别、分期等）	反映宏观状态	易获取，预后评估基础
基因突变	高维稀疏	驱动机制与靶点	指导精准治疗
拷贝数变异	中高维（大片段改变）	与进展 / 耐药相关	补充突变信息
基因表达	高维连续	反映功能状态	分型与疗效预测

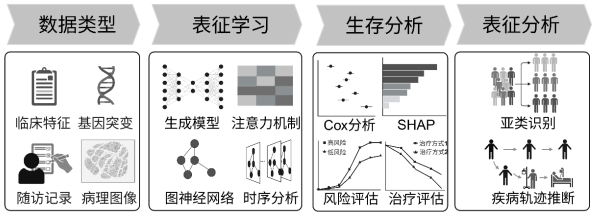

图 1 基于深度特征学习的肿瘤生存预测模型研究框架图

3 基于深度特征学习的肿瘤生存预测模型研究进展

肿瘤生存分析的核心目标在于通过解析复杂的临床与组学数据，实现对患者预后的精准评估与干预指导。传统的生存分析方法如 Cox 比例风险模型，主要依赖人工筛选的特征，难以捕捉高维数据中的非线性交互与深层生物学语义。随着深度学习技术的发展，研究范式已从单纯的统计回归转向基于深度表征的智能化全流程分析。下文将从患者特征有效表征、生存风险评估、表征分析三个层级，阐述基础模型如何解决临床核心问题。

3.1 基于深度学习的患者多模态特征有效表征

患者的高维生物学特征是生存分析的基石。临床多模态数据具有高维度、高噪声及稀疏性特点。如何从异质数据中提取出能够有效反映患者生物学本质的低维潜在表征（Latent Representation），成为构建基础模型的首要任务。Zhang 等提出了一种基于变分自编码器（Variational Autoencoder, VAE）的多任务深度学习框架 OmiEmbed，旨在解决多组学数据的维度灾难问题。该研究利用 VAE 将高维的基因表达、DNA 甲基化及拷贝数变异数据映射到统一的低维潜在空间，并通过多任务学习策略同时优化生存预测与表征重构损失。在 GDC 泛癌数据集上的实验结果显示，OmiEmbed 提取的深度特征在生存预测任务中取得了 0.772 的 C-index，显著优于原始特征及降维后的 PCA 特征。此外，Chen 等针对病理图像与基因组数据的融合问题，开发了 Pathomic Fusion 框架，利用卷积神经网络（Convolutional neural network, CNN）提取组织学图像的形态学注意力机制（Gated Attention Mechanism）实现了图像特征与分子特征的自适应融合，能够捕捉到肿瘤微环境中淋巴细胞浸润等细微的空间异质性。

3.2 基于深度特征的个体化精准生存风险预测

在获取有效的患者表征后，临床面临的最直接挑战是

对患者个体的生存风险进行精准量化。基于深度特征的生存预测模型通过引入非线性风险函数，显著提升了预测精度。Katzman 等提出了 DeepSurv 模型，通过多层感知机替代 Cox 模型中的线性风险函数，实现了对非线性协变量效应的建模。在 METABRIC 乳腺癌数据集的验证中，DeepSurv 的 C-index 达到 0.654。针对多模态数据的长期生存预测，Vale-Silva 等构建了端到端的 MultiSurv 框架，该模型能够直接处理全切片病理图像与临床数据，输出患者在不同时间节点的生存概率。结果显示，MultiSurv 在包含 33 种癌症的 TCGA 数据集上，其预测性能（C-index）相比仅使用临床特征的基线模型显著提升。Wen 等提出的 FGCNSurv 模型引入图卷积网络（Graph convolutional network, GCN）来建模患者间的关联，进一步提升了生存预测的鲁棒性。

3.3 相同风险人群的异质性驱动机制亚类识别

精准的风险预测仅解决了“预后如何”的问题，却未回答“为何如此”。临床实践中常出现“同分不同治”的困境：两组患者虽然具有相同的生存风险评分，但其背后的驱动机制截然不同，导致对同一治疗方案的响应存在巨大差异。因此，在风险评估的基础上，识别出具有明确生物学意义的隐匿亚类，是指导差异化治疗的关键。 Poirion 等提出的 DeepProg 框架，通过整合深度自动编码器（Autoencoder, AE）与 Cox 比例风险模型，不仅实现了生存预测，更通过无监督聚类在肝癌队列中识别出了具有显著预后差异的新亚型。这些亚型在传统 TNM 分期中无法区分，但深度特征揭示了其在特定信号通路上的异常激活。Rappoport 等则利用 SubtypeGAN 模型，通过生成对抗网络学习 Generative adversarial network, GAN）多组学数据的潜在分布，在乳腺癌中发现并定义了新的分子亚型。验证结果表明，这些基于深度特征聚类得到的亚型在生存曲线上表现出显著的分离度，且对应着完全不同的药物敏感性谱。

3.4 基于深度特征轨迹推断的疾病演化动态监测

肿瘤是一个动态演化的生态系统，患者的生物学状态会随时间推移发生克隆演化或耐药转型。即便明确了风险与亚型，缺乏对疾病演化轨迹的认知，临床医生仍难以把握最佳的干预时机。静态的生存预测模型无法回答“疾病目前处于哪个发展阶段”的问题，而基于深度特征的轨迹推断为解决这一时效性问题提供了新思路。借鉴单细胞生物学中的伪时间（Pseudotime）概念，研究者开始尝试在患者群体水平上构建疾病演化轨迹。Alaa 等提出了 Attentive State-Space 模型，该模型通过注意力机制对患者的纵向随访数据进行时序建模，不仅能够预测生存结局，还能推断出疾病潜在的演化路径和状态跃迁概率。在囊性纤维化与乳腺癌的队列研究中，该模型成功描绘了患者从稳定期向进展期恶化的风险流动轨迹。

4 面临的问题与挑战

尽管临床与多组学整合生存预测模型取得显著进展，但实际应用仍面临多方面挑战：（1）数据质量与异质性：多中心数据采集与测序平台差异导致批次效应显著，影响模型泛化。临床数据缺失普遍存在，虽然多重插补是常用策略，但在复杂多模态场景下，如何保证插补值的生物学真实性仍是难题 [4]。（2）样本量限制：高质量多组学数据获取成本高，导致样本量通常远小于特征维度。模拟研究指出现代深度建模技术对样本量需求更高，小样本高维场景易过拟合。（3）模型可解释性不足：深度模型黑箱特性影响临床可信度。Holzinger 等指出医学可解释（Explainable AI, XAI）需具备可验证性与可追溯性 [5]；Rudin 强调高风险决策应优先采用可解释模型或可解释框架。（4）外部验证与临床转化困难。（5）伦理与隐私问题：基因数据具有高度可识别性，即使去标识化仍存在再识别风险。

5 结语

临床与多组学数据整合的生存预测研究正从传统统计模型向深度学习与基础表征学习范式演进。深度模型通过表示学习、多任务训练与多模态融合提升了预后预测能力，并拓展至患者分型与疾病轨迹推断等任务。未来研究应重点关注：（1）构建多中心高质量标准化数据集；（2）发展更透明可解释的基础模型架构；（3）利用迁移学习与联邦学习解决小样本与隐私限制；（4）推进多模态融合与临床可行动性解释的统一框架。随着多组学数据积累与 AI 技术成熟，该领域有望在肿瘤精准医疗与个体化决策中发挥更大价值。

参考文献

[1] Sung H, Ferlay J, Siegel R L, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians, 2021, 71(3): 209-249.

[2] Bailey M H, Tokheim C, Porta-Pardo E, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell, 2018, 173(2): 371-385.e18.

[3] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798-1828.

[4] Gao J, Aksoy B A, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Science Signaling, 2013, 6(269): pl1.

[5] Holzinger A, Biemann C, Pattichis C S, et al. What do we need to build explainable AI systems for the medical domain? arXiv:1712.09923, 2017.