文章编号:1671-7554(2025)08-0017-24 DOI:10.6040/j.issn.1671-7554.0.2025.0512

多模态医学数据融合技术及应用

杨帆[1,2,3]

(1.山东大学齐鲁医学院公共卫生学院医学数据学系, 山东 济南 250012;
2.国家健康医疗大数据研究院, 山东 济南 250003; 3.山东大学齐鲁医院, 山东 济南 250012)

摘要:随着生物多组学、医学影像和电子健康记录等多源医疗数据的爆炸式增长,单一模态难以刻画复杂疾病的生物学异质性。多模态医学数据融合技术通过在特征级、表示级和决策级整合异构信息,为疾病预测与治疗提供了新的可能。本研究系统梳理了近年来基于深度学习与统计建模的融合方法学进展,包括Transformer与图神经网络驱动的端到端框架,贝叶斯及潜在因子模型支撑的显式概率推断,以及信息瓶颈、共性-特异性分解等增强表示有效性的理论新视角。针对跨模态异质性和高维稀疏性,本文总结了早期、中期、晚期三类融合策略及协同训练、多视角对齐等训练范式,并讨论注意力机制在捕获互补信息中的作用。进一步结合癌症预后、生物标志物发现、药物反应预测和临床决策支持等应用案例,阐释融合模型在提高预测性能、增强可解释性和契合临床工作流方面的优势与挑战。本文提出面向临床可落地的未来研究方向:构建安全合规的联邦数据湖、发展因果可解释融合框架、加强与医护流程的深度耦合,以实现从多模态数据到精准诊疗的闭环转化。

关键词:多模态融合;深度学习;信息瓶颈;可解释性;精准诊疗
中图分类号:R181.2+3 **文献标志码**:A

Multimodal medical data fusion technology and its application

YANG Fan[1,2,3]

(1. Department of Medical Dataology, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan 250012, Shandong, China; 2. National Institute of Health and Medical Big Data, Jinan 250003, Shandong, China; 3. Qilu Hospital of Shandong University, Jinan 250012, Shandong, China)

Abstract: With the explosive growth of multi-source medical data such as bio-multi-omics, medical imaging, and electronic health records, a single modality is unable to characterize the biological heterogeneity of complex diseases. Multimodal medical data fusion technology provides new possibilities for disease prediction and treatment by integrating heterogeneous information at the feature level, representation level, and decision level. This study systematically reviews the progress of fusion methodologies based on deep learning and statistical modeling in recent years, including end-to-end frameworks driven by Transformer and graph neural networks, explicit probabilistic inference supported by Bayesian and latent factor models, and new theoretical perspectives such as information bottlenecks and commonality-specificity decomposition to enhance representation effectiveness. In view of cross-modal heterogeneity and high-dimensional sparsity, this paper summarizes three types of fusion strategies, namely early, mid-, and late-stage, as well as training paradigms such as collaborative training and multi-view alignment, and discusses the role of attention mechanisms in capturing complementary information. Further combined with application cases such as cancer prognosis, biomarker discovery, drug response prediction, and clinical decision support, this paper explains the advantages and challenges of fusion models in improving prediction performance, enhancing interpretability, and fitting clinical workflows. This paper proposes future research directions for clinical implementation: building a secure and compliant federal data lake, developing a causal explainable fusion framework, and strengthening deep coupling with medical care processes to achieve a closed-loop transformation from multimodal data to precision diagnosis and treatment.

Key words: Multimodal fusion; Deep learning; Information bottleneck; Explainability; Precision diagnosis and treatment

收稿日期:2025-04-30
基金项目:国家自然科学基金(82273736)
通信作者:杨帆。E-mail:fanyang@sdu.edu.cn

随着医疗数据的多样化和大规模产生，如何有效融合多种模态数据以获得全面的临床洞察成为医学人工智能领域的关键挑战[1]。单一模态数据（如仅使用影像或仅使用基因组信息）难以全面反映复杂疾病的异质性，而临床医生在诊疗决策时通常会综合多种信息来源（如影像、实验室检查结果、患者病历等）[2-3]。多模态数据融合技术将不同来源的数据进行有机整合，以提高模型的预测性能和决策可靠性[3-4]。

近年来，面向医学多模态数据的融合技术取得了显著进展。一方面，方法学上涌现出大量基于深度学习的新模型（如 Transformer、图神经网络等）以及改进的统计融合方法（如贝叶斯融合模型、潜在因子模型、共性-特异性分解等），推动了融合算法性能的提升[3,5]。另一方面，数据层面越来越多类型的医疗数据被纳入融合范畴，包括生物多组学（基因组、转录组、表观遗传组等）、医学影像（放射影像和病理切片）、临床电子健康记录（electronic health record，EHR）以及患者可穿戴设备产生的连续监测数据[6]。针对不同模态数据的融合，研究者提出了多种融合策略（早期融合、中期融合、晚期融合）、训练范式（协同训练、多视角对齐、联合嵌入）和注意力机制，以充分挖掘各模态间的互补信息[1,3]。同时，融合模型的统计理论基础也受到重视，例如利用信息瓶颈原理提升表示有效性，以及增强模型可解释性以满足临床需求[7-8]。这些技术方法在癌症预后、生物标志物发现、疾病诊断分型、病理图像分析和临床决策支持等应用场景中展现出巨大潜力，相关研究成果层出不穷[9]。

多模态融合的核心驱动力在于这样一个基本假设：整合不同类型的数据能够产生协同效应，其整体价值大于各部分之和[10]。不同数据模态捕捉了生物系统在不同尺度、不同层面的信息，这些信息之间可能存在复杂的相互作用和关联。有效的融合方法不仅是简单的数据拼接，更需要能够揭示并利用这些跨模态关联，从而构建更接近生物学真实情况的模型。因此，对融合方法学的深入研究，探索如何最有效地提取和整合这些协同信息，是实现多模态数据价值的关键[11]。

此外，许多研究指出，开发基于人工智能的多模态融合模型，其目标之一是模拟临床医生在诊断和决策过程中整合多种信息来源（病史、体征、检查结果、影像报告等）的思维过程[11]。临床实践本身就是一种多模态信息融合的过程。这种与临床工作流的对齐，不仅要求融合模型具有高预测精度，还对其可解释性提出了内在要求。一个能够提供决策依据且易于理解的模型，更容易被临床医生信任和接受，从而促进其在实际医疗场景中的应用[12]。因此，融合方法学的发展不仅要追求性能的提升，还需要关注模型的可解释性和与临床实践的契合度。

需要说明的是，以往已有学者对多模态融合技术进行了总体性综述[13-15]。然而，医学领域具有一些特殊需求（如数据稀缺[16]、严格的可靠性和隐私要求等）。本文在前人工作的基础上，突出医疗应用场景的特点，对关键方法进行深入分析和形式化描述，并通过横向比较不同方法的优缺点，提炼出医疗多模态融合所面临的独特挑战与未来机遇。

1 多模态数据融合方法

多模态融合方法学大体可分为基于深度学习的方法和统计建模的方法两大类，同时近年来还出现了融合先验知识的图模型方法。本节分别介绍各类方法的代表性进展，并对其算法特点进行比较。

1.1 深度学习驱动的多模态融合

深度学习以其强大的非线性建模能力，成为近年多模态数据融合的主流方法。相比传统方法，深度学习能够自动学习跨模态的复杂关系和高层语义表示[17]。常见的深度学习融合架构包括全连接神经网络、卷积神经网络（convolutional neural network，CNN）、循环神经网络（recurrent neural network，RNN）、自编码器（autoencoder，AE）等，以及近年来兴起的 Transformer 和图神经网络（graph neural network，GNN）等新型架构[18]。下面重点讨论 Transformer 系列模型、GNN 和多模态自监督学习（self-supervised learning，SSL）在医学数据融合中的应用（图 1）。

1.1.1 Transformer 融合模型

Transformer 以其自注意力机制在自然语言处理领域取得成功，近年来被广泛引入多模态融合任务中[3]。Transformer 架构通过多头注意力实现不同模态特征的交互建模，能够在融合过程中自动关注重要的模态信息[19]。典型做法是将各模态的特征表示映射为序列化的 token 序列，再送入Transformer 编码器/解码器进行融合。这样，模型可以在每一层注意力中动态调整各模态特征的重要性，从而学习到统一的跨模态表示。

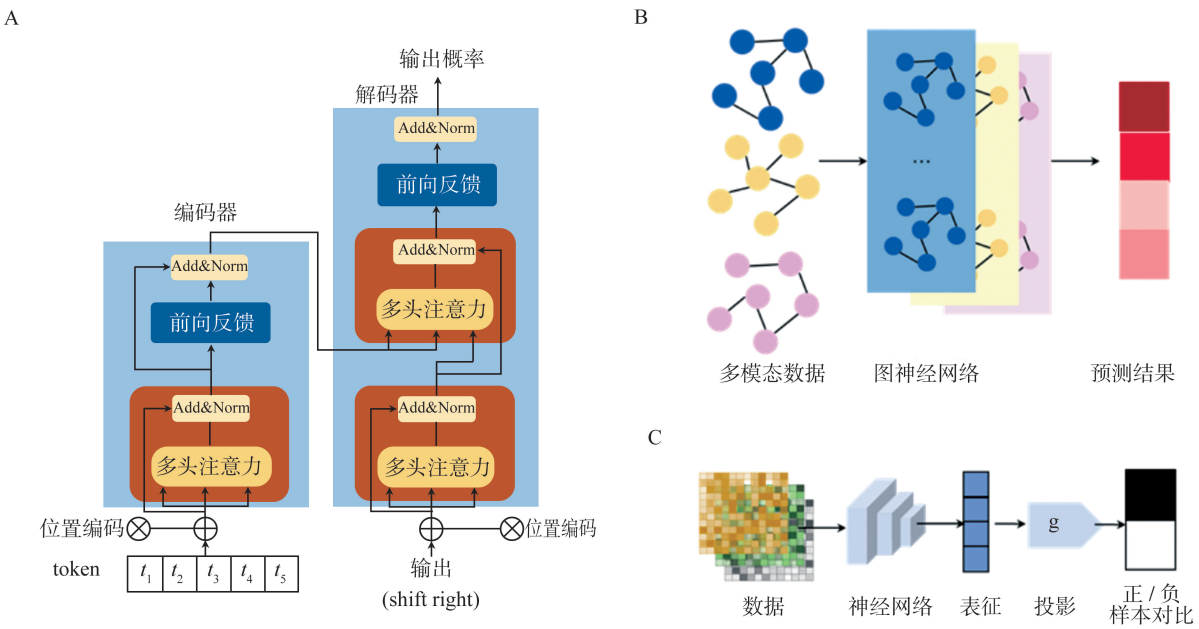

图 1　基于深度学习驱动的多模态融合

A：Transformer 融合模型示意图；B：图神经网络融合示意；C：多模态自监督预训练示意。

Figure 1　Deep learning-driven multimodal fusion

A: An illustration of a Transformer-based fusion model; B: Graph neural network-based fusion; C: Multimodal self-supervised pretraining schematic.

多项研究已经证明，基于 Transformer 的融合在医学任务上优于传统融合方法[20]。例如，Zhou 等[21]构建了统一的 Transformer 模型，将胸部 X 光影像与实验室检测数值数据融合，用一致的 token 序列表示影像和临床信息，在辅助诊断中取得了优于经典早期/晚期融合的效果。Nguyen 等[22]提出了多 Agent Transformer 架构，为每种模态（如 MRI 影像和临床指标）设计单独的 Transformer 分支，再通过交互注意力机制融合不同模态以预测阿尔茨海默症进展，性能超过以往的融合模型。还有 Khader 等[23]针对 ICU 重症监护数据开发了大规模 Transformer 双编码架构：一个 Transformer 编码器提取临床数值和影像特征，另一个 Transformer 将二者融合用于患者生存预测，证明该模型可高效扩展到大型数据集。值得一提的是，Transformer 的注意力权重还天然提供了可解释性：模型可以自适应地权衡不同模态的重要程度，定位诊断决策所依赖的关键特征[19]。总的来说，Transformer 通过注意力机制实现了对多模态数据的深层次融合，已在医学影像+临床数据融合[7]、多模态时间序列预测等任务中取得显著进展。然而，由于 Transformer 模型参数量大、对数据规模要求高，在小样本医疗数据场景下可能出现过拟合，需要结合迁移学习或预训练技巧加以克服。

1.1.2　GNN 与多模态图表示

GNN 可以在图结构数据上执行深度学习，其兴起为结合生物医学先验知识与多组学数据提供了新途径[24]。在多模态融合中，GNN 常用于表达模态内部或模态间的关系，例如，将基因、蛋白等构成生物分子相互作用网络，将影像中不同区域作为结点构建拓扑图，或者构建患者-特征二部图来整合不同类型的数据。通过在图上应用卷积或注意力操作，GNN 能够捕捉模态内的局部模式及模态间的拓扑关联，从而实现信息融合[25]。

在多组学数据融合方面，一项典型工作是 MOGONET[5]。该方法为每种“omics”数据（如 mRNA 表达、DNA 甲基化、miRNA 表达等）构建样本相似图，用图卷积网络（graph convolutional network，GCN）提取每种组学的判别特征，然后设计视图关联发现网络（view correlation discovery network，VCDN）在标签空间融合各组学的初步预测。通过联合训练，MOGONET 能够同时学习组学内模式和组学间的相关性，分类性能优于传统多组学集成方法，并能识别各组学中重要的生物标志物。

在影像与图形数据融合方面，GNN 也展现出独特优势。例如，将病理全视野数字切片（whole slide image，WSI）划分为众多区域块并构建空间邻接图，再将每个块的视觉特征作为结点嵌入，同时引入

该病例的基因表达数据，可以利用图注意力网络融合空间图像特征与全局分子信息用于预后预测[26]。Zheng等[27]提出的模型中，每张病理图被表示为结点（代表组织区域）和边（代表空间邻近关系）的图结构，Bulk测序得到的基因表达向量通过全局注意力模块融合进图的结点表示中，然后经过若干层“图混合”网络（包括结点混合和通道混合），最后通过注意力池化和全连接层预测患者生存风险。这种方法有效捕捉了基因表达与组织形态之间的对应关系，在多种癌症的预后预测中表现出色。总之，GNN类方法善于融合同一实体的不同模态信息（如同一患者的多组学或同一组织的影像与分子数据），通过图结构将先验知识（如生物网络、空间邻近等）显式引入融合，有助于提高融合模型的生物学合理性和可解释性。需要注意的是，GNN模型通常包含较多超参数（如图构建方式、卷积层数等），对数据质量和先验网络依赖较大，因此在应用时需谨慎进行图构建和模型调优，以防噪声关系干扰模型学习。

1.1.3 SSL与预训练

由于有标签的医学数据匮乏，SSL在医学领域备受关注[21]。SSL通过设计预训练任务来挖掘未标注数据的跨模态关联，为下游有监督任务提供良好的初始化或表示[28]。常见的多模态预训练策略包括对比学习、重构预测（如跨模态重构或掩码建模）以及多任务预训练等。

在医疗场景中，一个经典案例是影像-文本模态的对比学习。Zhang等[29]提出的ConVIRT方法利用大量未标注的放射影像及对应报告，通过图像和文本编码器以及对比损失函数，使得正确匹配的影像-报告对在嵌入空间距离更近，不匹配的对更远。这种预训练让影像编码器学到了与临床语义相关的特征，在下游如X光影像疾病分类等任务上显著提升了性能。后续还有MedCLIP等改进方法进一步考虑非成对数据的对比训练，扩大了可利用的预训练数据范围[30]。除了影像-文本，对比学习也用于生理信号与症状文本、视频与传感器数据等多种模态组合，在异常检测和事件预测中取得进展。

另一类SSL是跨模态重构，例如训练模型从一种模态生成或重构另一种模态的数据。医疗中的应用包括：利用多模态变分自编码器（variational autoencoder，VAE）同时对影像和非影像数据进行编码-解码，在压缩表示中迫使模型提取两种模态的共同信息[31]；或者让模型学会从心电图信号重构对应的脉搏波（photoplethysmography，PPG）信号，迫使其提取心血管状态的公共特征，用于预后评估。这类重构任务利用不同模态之间的关联来学习联合表示。此外，还有研究设计拼图解谜、多模态排序等预训练任务，让模型预测打乱的多模态数据的正确排列，以学习跨模态的全局一致性。

自监督预训练可极大缓解多模态融合对大样本数据的需求[32]。预训练得到的模型在下游小样本任务上往往表现出更好的泛化性能。例如，Ghassemi等[33]利用自监督学习将EHR时间序列数据和临床文本预训练为统一表示，在迁移到ICU患者预后预测任务时，取得了比从零开始训练更高的AUC。此外，自监督还能实现模态表示对齐，预训练后的各模态嵌入通常位于一个联合的表示空间，为后续简单融合（如特征拼接）奠定了基础。

未来，随着医学领域大模型的发展，基于海量多模态数据预训练的通用模型（如多模态医学大语言模型）有望实现“一模多能”，在少样本甚至零样本情况下，为各类任务提供强大支持[34]。当前仍需解决的问题包括：如何设计更符合医学语义的预训练任务，如何融合更多类型的数据源，以及如何平衡规模庞大的多模态预训练模型与有限的医疗数据之间的矛盾等。

1.2 统计融合模型

在深度学习大放异彩的同时，传统统计模型和概率图模型仍然是多模态融合的重要工具，尤其在样本量有限、需要可解释性和不确定性量化的场景下。此类方法通常假设数据由潜在的统计结构生成，通过建模不同模态的数据分布及其关联来实现融合[35-36]。本节介绍三类具有代表性的统计方法：贝叶斯融合、潜变量模型以及共性-特异性分解模型。

1.2.1 贝叶斯融合方法

贝叶斯方法通过显式地定义概率模型，将多模态融合问题转化为后验推断，能够天然处理不确定性并融合先验知识[35]。在多模态融合中，一个基本思想是构建贝叶斯层级模型，假定存在某些潜在变量影响各模态数据的生成，从而在贝叶斯框架下联合推断这些潜在因素及感兴趣的结果。典型例子包括贝叶斯多组学因子分析、贝叶斯网络等。

1.2.1.1 贝叶斯因子模型

Samorodnitsky等[37]提出了贝叶斯同时因子分解（Bayesian simultaneous factorization，BSF）的方法，将多组学数据分解为共享因子和特异因子，同时结合对临床表型的预测建模。该模型在贝叶斯框架下引入先验正则，利用Gibbs抽样或变分推断进行

求解,使得推断得到的因子自带不确定性估计和稀疏性。相比传统矩阵分解,贝叶斯方法有效避免过拟合,并可自动选择因子数。类似地,Ghosal 等[38]应用贝叶斯多视图判别分析整合基因组和影像数据,将判别任务融入因子提取过程中,使得到的低维表示对预测任务更有用。这些贝叶斯模型在癌症亚型识别、复杂疾病预测中取得了比简单融合更好的效果,同时提供了因子层面的生物学解释[31]。

1.2.1.2 贝叶斯网络(Bayesian network, BN)模型

BN 以有向无环图表示变量之间的条件独立关系,可用于描述多模态数据的因果或关联结构。在融合中,可以构建一个 BN,其中节点包括不同模态的特征(甚至潜在变量)以及最终的疾病表型,通过学习 BN 的结构和参数,实现模态间关系的融合推理。例如,Suter 等[36]提出的 bnClustOmics 方法将多组学数据的融合聚类建模为“贝叶斯网络混合模型”。该模型假设存在若干潜在簇,每个簇对应一个 BN,网络节点涵盖各组学的特征。通过对癌症多组学数据拟合 bnClustOmics,不仅可发现患者的聚类亚型,还能得到每个亚型对应的 BN 结构,揭示基因-蛋白-表型之间的相互作用机制。在对肝癌数据的应用中,该方法识别出了三个主要癌症分子亚型,并找出了不同亚型中特有的分子网络,其中部分网络特征与患者生存显著相关。这表明 BN 能将融合聚类和机制发现有机结合,为多模态数据提供更深入的解释。除了 BN,也有研究采用马尔可夫随机场等无向图模型,将不同模态的变量看作共同的随机场,通过估计图结构和相互作用参数实现信息融合。例如,使用马尔可夫网络将影像特征与临床变量连接,通过能量函数将它们与疾病风险的关系进行建模[39]。总体而言,贝叶斯融合方法提供了概率推理视角,可以有效结合先验知识和不确定性处理,但往往计算代价较高(需要采样或近似推断),对于高维数据可能需要借助降维或稀疏先验来提高可行性。

1.2.2 潜变量模型与多模态因子分析

潜变量模型假设观测的多模态数据由低维的潜在因子生成,各模态可能共享某些因子但也存在各自特有的因子。通过推断这些潜在因子,可以在融合中提取出模态间的共性信号和特异模式[40]。这类方法包括经典的主成分分析(principal component analysis, PCA)扩展、多视图 CCA、PLS,以及近年来发展的多组学联合因子分析等。

一个代表性模型是多组学因子分析(multi-omics factor analysis, MOFA)及其改进版 MOFA+[41]。MOFA定义每个组学数据矩阵由一组共享潜在因子和组学特异因子线性生成,并在变分贝叶斯框架下估计因子载荷和因子值。从形式化角度,假设每个模态的数据矩阵 $\boldsymbol{X}^{(m)}$ 可近似分解为:

$$\boldsymbol{X}^{(m)} \approx \boldsymbol{Z}\boldsymbol{W}^{(m)} + \boldsymbol{E}^{(m)},$$

其中,$\boldsymbol{Z}$ 为共享的潜在因子矩阵,$\boldsymbol{W}^{(m)}$ 为模态 m 对应的因子载荷矩阵,$\boldsymbol{E}^{(m)}$ 为噪声项。通过变分推断可同时估计出共享因子和各模态特有因子在每个样本中的取值。结果是每个潜在因子都具有一定的生物学意义,可能对应某种分子通路或细胞状态,对所有组学都有贡献;同时各组学的特异因子解释每种数据中独有的变化。MOFA 在多个癌症数据集上成功识别出了已知的驱动分子模式(如特定基因突变状态作为一个因子影响基因表达和表观遗传)。由于 MOFA 采用概率生成模型,还可处理缺失模态的数据,具有一定鲁棒性。具有类似思想的还有 iCluster 系列方法,其将多源数据联合聚类为若干亚型,同时提取共同的主成分用于可视化和下游分析[42]。近年来,不少深度生成模型也被用于潜变量融合,如 VAE 和生成对抗网络(generative adversarial network, GAN)等[31]。

潜变量模型的优势在于实现降维融合:它将高维异质数据映射到共同的低维子空间,去除了噪声和冗余,保留了与主要生物学信号相关的部分[43]。这些低维因子常常有助于理解模态间关系。例如,在阿尔茨海默症研究中,一个共享因子可能同时捕捉到 MRI 影像的脑萎缩模式和血液转录组的炎症基因表达升高,从而将这两个模态关联到疾病状态[44]。综上,潜变量方法提供了一种紧凑表示多模态数据的途径,使融合模型更易于训练和解释。然而,需要注意潜在因子数量的选择以及潜在语义的可解释性。一些最新研究探索了稀疏因子载荷和旋转不变性技术,以获得更具生物学意义的因子。

1.2.3 共性-特异性分解模型

共性-特异性分解旨在将多模态数据分解为共享成分和模态特有成分。这是一种特殊的潜变量模型,更加强调区分各模态的共同信息与独有信息。早期的代表方法是联合与个体变异分解(joint and individual variation explained, JIVE)[45]。它通过 PCA 分别求解多个数据块的共同低维子空间和各数据块的独立子空间,使每个模态的数据表示为这两个子空间之和。JIVE 能够揭示不同数据源之间的关联结构,并量化每个模态中无法由其他模态解释的变异。近年又发展出许多改进版本,如基于角度的联合与个体变异分解(angle-based JIVE,

AJIVE）将稳健性和统计检验引入分解过程[46]。

在 2020 年后，不少研究将共性-特异性分解思想与贝叶斯方法相结合，提升其灵活性和预测能力。例如，上文提到的 BSF 模型可视作贝叶斯版本的共性-特异性分解[47]：它同时分解多组学数据和相关的表型，使数据的联合部分和特异部分都服务于表型预测，从而实现融合分析和预测的统一。另一些工作将共性-特异性思想用于特定任务：如 Yang 等[48]将患者的多模态特征分解为疾病相关（共享）和个体背景（特异）两部分，仅用共享部分进行预测以降低协变量偏倚的影响，取得更稳健的结果。还有研究在深度学习模型中显式加入“共享/特异”分支，例如训练两个子网络分别提取跨模态共有特征和模态私有特征，并对共有特征施加一致性约束、对私有特征施加去相关约束，以提高融合效果和可解释性。

从形式上看，共性-特异性分解通过明确公共和特异信息建模，实现了信息模块化。可利用数学公式将每个模态的数据矩阵表示为：

$$\boldsymbol{X}^{(m)} = \boldsymbol{J} + \boldsymbol{A}(m),\ \boldsymbol{X}^{(m)} = \boldsymbol{J} + \boldsymbol{A}^{(m)},$$

其中，$\boldsymbol{J}$ 表示所有模态共享的部分（位于公共子空间），$\boldsymbol{A}^{(m)}$ 表示模态独有的部分（位于模态特有子空间）。通过约束 $\boldsymbol{J}$ 和各 $\boldsymbol{A}^{(m)}$ 在子空间上的正交性，可确保提取的共享信息和特异信息相互独立[45]。共性-特异性分解的优势在于清晰的模块化，能够明确地区分哪些信息是所有模态共有的（可能代表潜在的疾病状态或总体趋势），哪些信息是某一模态独有的（可能代表模态特定噪声或特殊视角）[49]。这种区分在医学上很有意义。例如，在多组学癌症数据中，共享成分可能对应肿瘤的主要分子亚型，而特异成分则包括测序平台特有的批次效应或组织来源的差异。基于共享成分的疾病诊断分类可提升模型稳定性与泛化能力，因其主动隔离了特异性噪声。

需要注意的是，如果各模态之间相关性本身不高，强行提取共享部分可能遗漏有用信息。因此实践中需要根据模态关联程度调整模型，例如对于弱相关的多模态数据，可降低共享成分的维度，让模型更多依赖各模态的特异成分分别学习，再在决策层融合。

1.3 图结构融合方法

“图模型”在本文中特指以图结构来表达多模态数据关系的融合方法，包括但不限于概率图模型（如上文讨论的 BN）以及各类知识图谱、关系网络等。与第 1.2 节偏重概率推断的视角不同，本部分强调利用“显式的图结构”来融合和分析多模态数据的关联。这方面的研究通常将多模态数据中的实体和特征用节点和边表示，然后在图上应用算法（图搜索、网络分析、传播等）实现信息整合（图 2）。

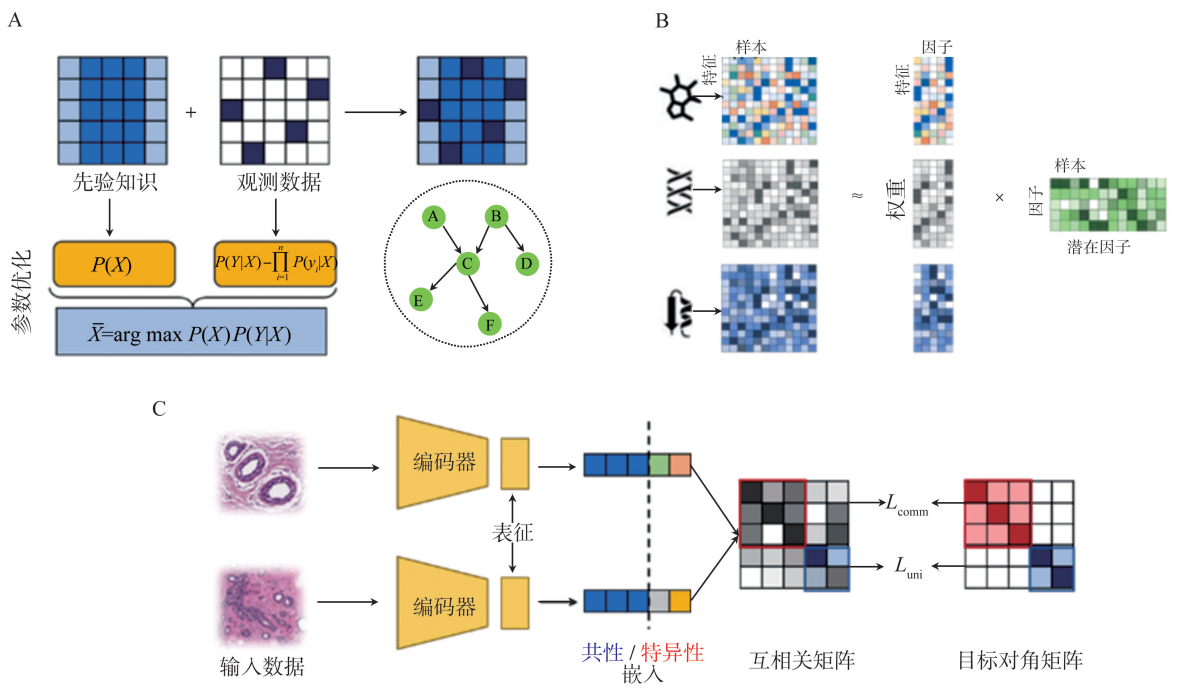

图 2 基于统计模型驱动的多模态融合

A：贝叶斯融合方法；B：潜变量模型与多模态因子分析；C：共性-特异性分解模型。

Figure 2 Statistical model-driven multimodal fusion

A: Bayesian fusion methods; B: Latent variable models and multimodal factor analysis; C: Joint-and-individual variation models.

图模型应用的一个重要方向是医学知识图谱与推理。科研人员构建了大型医疗知识图谱,将基因、疾病、药物、症状等实体通过已知关系连接成网络。如果研究者有患者的多模态数据(如基因突变列表、临床症状文本),就可以将这些数据映射到知识图谱上,再通过图中的关联边来传播信息或寻找连接路径(图 3)。例如,某患者的基因 A 出现突变,而知识图谱中基因 A 通过一系列关系指向疾病 X,则结合患者的其他检查数据,可以推测该患者患疾病 X 的风险[50]。又如,将影像中检测出的病灶与知识图谱中对应的解剖部位节点相连,再根据该部位相关的临床事件节点,辅助进行诊断决策。这类融合更像一种推理过程,常用的技术包括路径搜索、随机游走嵌入,以及图注意力网络(这部分方法与前述 GNN 有所重叠)。近年来,一些工作将知识图谱应用于药物反应预测,融合化合物结构、基因表达和表型数据,通过图中的关联推荐潜在治疗方案[51]。知识图谱方法的优势是能引入海量的生物医学背景知识,提高预测的可靠性和可解释性,但需要大量专家构建和维护高质量的医学图谱。

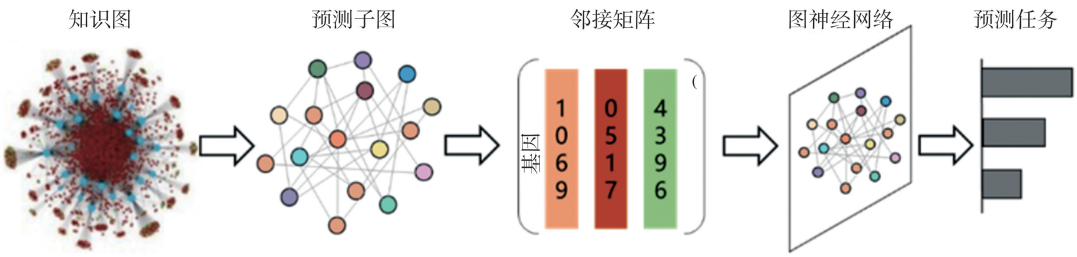

图 3 基于图结构模型驱动的多模态数据融合
Figure 3 Graph-structure-driven multimodal data fusion

另一个图模型的应用是多模态实体对齐与链接。例如,在数字病理中,将同一肿瘤患者的病理图像和对应的基因组数据通过共同的解剖结构进行对齐,把病理图像划分的区域与特定基因表达的空间分布联系起来,从而构建空间分子图谱。这种方法在单细胞多模态测序和空间转录组技术中非常流行:同一个细胞既有形态图像又有多组学数据,可构建一个细胞间相似网络或细胞-基因二部图,然后用图算法(如社区发现或标签传播)来发现细胞异质性或分子模式[27]。再如,可穿戴设备记录的日常活动数据可以和临床事件用时间轴对齐,构成时间序列因果图,通过图中的路径强度来衡量不同事件的关联,以调整模型中的协变量偏倚[39]。这些图模型手段强调对不同模态数据间关联关系的显式建模,使融合不局限于特征层面的拼接,也体现在关系层面的推理。

需要注意的是,图结构融合方法往往需要预先定义图结构或使用现有知识,这对数据融合提出了不同于端到端学习的要求。在实际应用中,常将图模型与机器学习结合,如图正则化在训练融合模型时加入一项损失,鼓励模型的输出与先验知识图中已知的关系一致,以提升模型的可信度;又如图嵌入融合先对知识图谱训练,得到节点的嵌入表示,然后将这些嵌入作为额外特征与原数据一起输入模型,以提供背景支持。总之,图结构方法为融合提供了另一个视角,即利用网络拓扑信息来增强多模态集成。这在医疗 AI 迈向知识驱动和因果推理的过程中将扮演重要角色。

1.4 各类方法的比较与分析

综上所述,不同类别的多模态融合方法各有特点,在应用条件和效果上有所差异,见表 1。深度学习方法擅长从大规模数据中自动学习复杂的非线性模式,对数据规模要求高,但对小样本问题可能不够稳健,需要结合预训练或迁移学习技术。其可解释性相对较弱,不过诸如注意力机制等设计在一定程度上提供了直观解释。统计模型方法具有良好的理论可解释性和不确定性量化能力,适合小数据情境,可融合先验知识避免过拟合,但通常依赖模型假设(如线性、正态等),对高度非线性的关系拟合不足,模型扩展性和灵活性也弱于深度学习。图结构融合方法能够引入显式的领域知识和关联关系,使模型预测更具可解释性和可信度,但往往需要高质量的知识图谱或预定义关系网,构建成本高,且如果知识不完备可能限制模型表现。与此同时,图模型与深度学习、统计方法并不矛盾,反而可以结合使用,例如在深度模型训练中加入图正则约束,在统计模型中融入网络先验等,从而兼顾自动学习能力和先验知识。

表1 不同多模态融合方法类别的特点

Table 1 Characteristics of different multimodal fusion methods categories

方法类别	优势特点	局限与挑战	典型应用场景
深度学习方法	自动学习复杂非线性关系;可端到端优化;性能随数据量增加而提升	需要大量标注数据;可解释性较弱;小样本时易过拟合	大规模多模态数据融合;需要高预测精度的任务
统计模型方法	理论基础成熟(概率框架);结果可解释(因子/网络解释);处理不确定性	模型假设限制(线性等);难捕捉复杂非线性;计算成本高	样本少但需可靠性的任务;注重机制发现和假设检验的研究
图结构融合方法	融入领域知识(知识图谱等);提高可解释性和可信度;关系推理能力	依赖高质量先验知识;构建维护复杂;对未知关系缺乏适应性	需要结合生物知识的场景;强调因果关系和推理的任务

总体来说,对于数据充足、模式复杂的问题,深度学习方法往往是首选;当数据有限或强调机制解释时,统计建模和图结构方法更具优势;在需要强鲁棒性和可靠性的医疗场景,可以考虑将二者结合,构建融合先验知识的深度模型。未来的发展趋势可能是多种范式的融合,例如借助因果推断理论指导深度模型设计,或开发同时包含神经网络组件和概率图组件的混合模型,从而扬长避短,充分发挥协同效应。

通过对不同方法的横向比较可以发现,没有单一的融合策略适用于所有问题。研究者需要根据具体任务和数据特点选择或设计合适的融合方法。在实践中,也可以采用集成融合的思路,同时训练多种类型模型,取其优势互补。将深度模型的高准确率结果与统计模型的不确定性评估结合,或者在深度框架中嵌入图结构先验,这将有助于在提升性能的同时,确保模型输出的可靠性与可解释性。

2 多模态数据类型

医疗领域涉及的数据类型繁多,不同模态的数据在尺度、维度、采样频率、信息含义等方面差异巨大[16]。常见的医学多模态数据可大致分为以下几类:① 生物多组学数据;② 医学影像数据(含放射影像和病理影像);③ 临床电子病历数据;④ 生成的患者健康数据(如可穿戴设备记录)。本节将逐一介绍这些数据类型及其融合特点。

2.1 生物多组学数据

生物多组学指来自生物体不同分子层次的大规模数据,包括基因组(DNA 序列变异)、转录组(mRNA 表达)、表观基因组(DNA 甲基化和组蛋白修饰)、蛋白质组、代谢组等(图4)。高通量技术的发展使得研究者能够同时测定同一批患者的多种组学数据,为疾病机制和精准医学研究提供全景式视角[52]。不同组学捕获了生物系统在不同层面的信息,例如 DNA 突变揭示遗传驱动因素,RNA 表达反映基因活性状态,甲基化代表基因调控情况,蛋白质和代谢物则体现细胞和代谢通路的功能状态。如何融合分析这些组学数据以挖掘它们之间的关联,是多组学数据融合的核心问题之一。

多组学数据融合的一大挑战是高维且异质。每种组学的特征数量都可能远超样本数(如基因表达矩阵常为上万维,而代谢物仅上百维),且不同组学的数据动态范围、噪声分布往往不同。另外,不同组学之间存在复杂的相依关系(如 DNA 上的突变会影响 RNA 的表达)。为此,研究者常用降维和统计建模方法(参见1.2节)来融合多组学。例如,通过共性-特异性分解提取共同因子以关联不同组学的变化[53];或使用图模型引入生物网络(如蛋白质相互作用网络)帮助在多个组学间传播信息[5]。在应用上,多组学融合已用于癌症预后、生存分析、药物反应预测(整合基因组和转录组寻找药物敏感性标志)、复杂疾病的分型(如利用基因+代谢组评估心血管疾病风险)等。总的趋势是从单一组学的分析走向全景多组学的集成,这也催生了如 TCGA 等大规模项目,其收集并公开了配对的多组学-临床数据资源。然而,多组学数据常存在缺失模态的情况,并非每个患者都有所有类型的测定。针对这一问题,融合方法需要具备处理缺失数据的机制,如采用矩阵补全、模型推断缺失值或仅使用有公共模态的子集数据训练等[54]。尽管存在这些挑战,多组学融合已被证明能够提供比单组学分析更稳健且有洞察力的结果,为精准医疗提供新工具。例如,在植物生物学研究中,类似的多通路整合调控机制也展现出系统层面的复杂性,这种多因子协同作用的通路调控模式与多组学融合分析思路高度一致,均强调在系统水平上整合多种信号与数据层次,从而更全面精准地解析生物过程[55]。

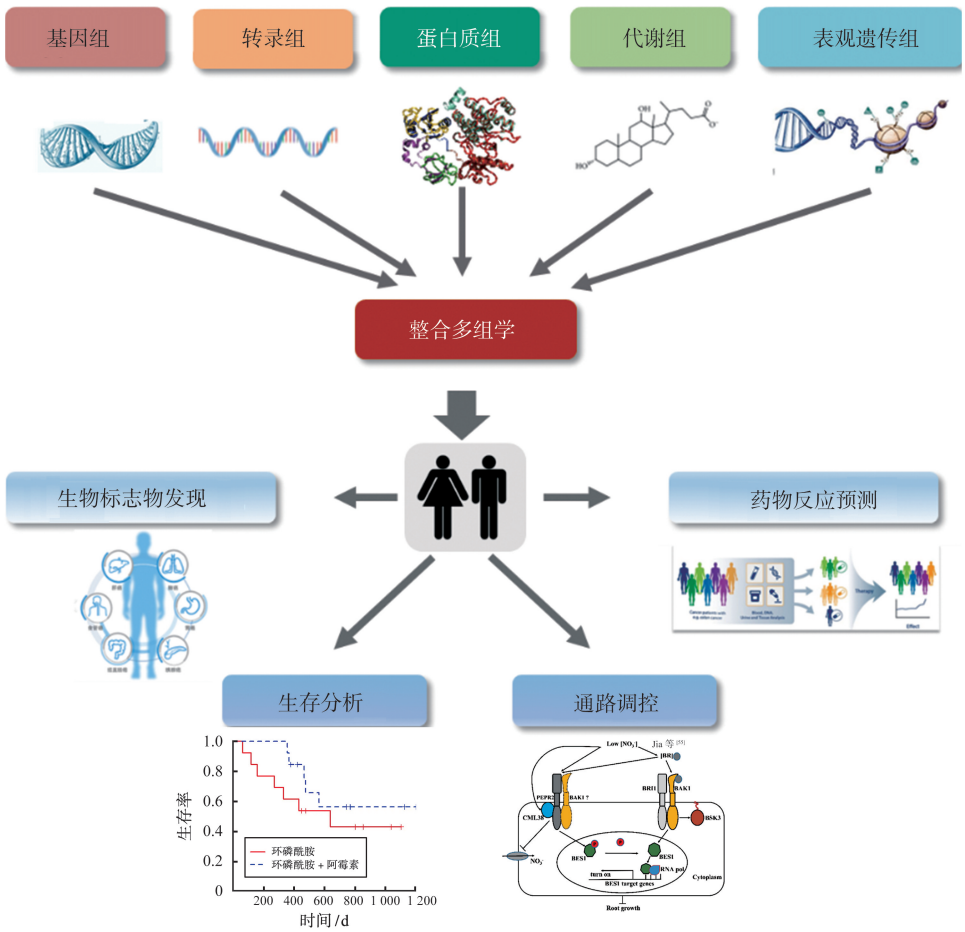

图 4　面向生物组学的多模态组学数据

Figure 4　Multimodal omics data for bioinformatics

2.2　医学影像数据(含放射影像组学)

医学影像是医疗诊断的重要数据模态,包括放射影像(X 光、CT、MRI、超声等)和功能分子影像(PET 等)。此外,影像组学指从医学影像中提取定量特征,用于建模和决策支持[56]。单一的影像模态已经提供了丰富的信息,例如 CT 可见解剖结构,PET 反映代谢活性等。然而,融合多种成像模态,或者将影像与其他类型的数据结合,往往能够显著提高诊断准确性并加深对疾病的理解。

在临床实践中,多模态影像融合最常见的情况是多种影像模式的结合。例如,对于疑难肿瘤的诊断,医生常同时参考 MRI 和 PET 影像:MRI 提供高分辨率的软组织解剖,PET 揭示肿瘤的代谢热点。将两种影像通过配准叠加,可以同时看到肿瘤的位置和活性,实现更准确的分期和疗效评估[57]。在人工智能模型中,也有融合多种影像的算法研究。例如,将 CT 和 PET 的特征图在 CNN 中按通道维度拼接(即早期融合),用于肺结节良恶性的判断,效果优于单一模态模型[58]。又如序列影像的融合:在心脏成像中融合解剖结构影像与功能影像,或者在神经成像中融合结构 MRI 和功能 MRI,从而同时捕获静态解剖和动态活动信息,用于疾病预测。

除了多种影像模态本身的融合,将影像与其他模态结合也是当前研究的重要方向。例如“影像+临床数据”将患者的临床变量(年龄、性别、实验室结果等)与影像一起用于模型训练,往往能提高预测性能[59]。Li 等[60]利用深度学习融合视网膜眼底影像和临床危险因素来预测糖尿病发生,结果发现融合模型的准确率显著高于仅用影像或仅用临床数据。这说明医学影像与其他模态的数据具有很强的互补性,即影像提供形态学证据,临床数据提供生理/病理学背景,两者结合能更全面地刻画疾病状态。

影像数据融合的另一个特点是可以采用传统图像处理和计算机视觉的方法。在多模态融合模型中,有时需要先对影像进行预处理(如配准、分割)以对齐其他模态的数据。例如,将病灶区域从影像中分割出来,再与对应的基因检测结果匹配,这样融合时模型能聚焦于影像中的相关区域,提高效率[61]。Radiomics 特征提取也是常用手段,通过固定算法从影像中提取数百上千维的定量特征(如纹理、形状、灰度直方图等),然后将其与基因表达或

临床变量一起输入机器学习模型[62]。Radiomics 技术将高维影像转换为结构化的数据特征，使之更容易与其他表格型数据进行融合分析[63]。Zheng 等[64]提取 MRI 肿瘤纹理特征并与患者的基因型融合建模，发现某些纹理模式和特定基因突变的组合对疗效具有预测价值。

总体来说，医学影像在多模态融合中扮演着“视觉主角”的角色，常需要与其他“非视觉”模态结合以丰富诊断信息。未来，随着多模态成像技术的发展（如 PET/MRI 一体机、全息全景成像等）以及影像与组学、文本融合方法的成熟，有望看到医学影像融合应用于更加精细的疾病分型和疗效预测。在应用中也需注意，影像数据量巨大，融合时的计算成本和存储开销较高；不同影像模态之间还存在配准误差等实际挑战，需要通过算法和工程手段加以克服。

2.3 数字病理数据

数字病理指将病理组织切片扫描成超高分辨率 WSI，用于计算机辅助分析。病理图像包含微观层面的组织结构和细胞形态信息，被视为诊断的“金标准”数据。面向病理组学的多模态数据（包括病理影像、临床信息、基因数据等）将病理图像与其他分子或临床数据融合，可以深入理解疾病的组织学-分子学关联[65]（图 5）。

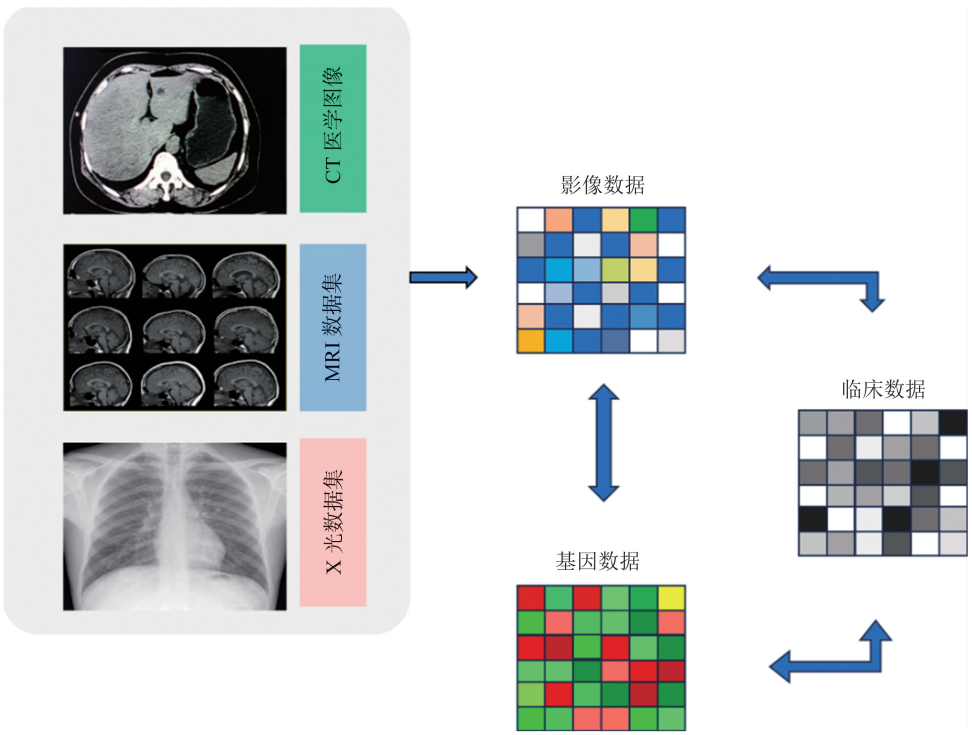

图 5 面向病理组学的模态数据

Figure 5 Modal data for pathomics

病理图像的特点是尺寸巨大（单张全切片图像可达数 GB 像素）且信息极其丰富（细胞形态、基质特征、免疫浸润等）。直接将病理图像与其他数据融合在技术上有难度，因此大多数工作采用“特征提取+融合”的范式。一种做法是通过弱监督学习从 WSI 中提取全局表征向量，然后与基因组等数据拼接融合（图 5）。Chen 等[66]提出了弱监督深度网络，从每张 WSI 中自动学习一个能预测生存的特征向量，同时用 DNA/RNA 数据训练另一个网络，然后在预测层进行晚期融合，实现对多癌种患者的生存预测。该研究发现融合模型较只用病理或只用基因数据的模型有更好的 C 指数，证明了两种模态的互补性。

另一些研究尝试更紧密的融合，例如利用病理图像的空间信息。Brussee 等[67]提出 GraphS-GAN 框架，将病理 WSI 划分为图像块并构建空间邻接图，引入基因表达数据作为附加节点或节点特征，通过图注意力网络实现早期融合，在乳腺癌和肺癌数据上提高了生存预测的准确率。该方法还能输出图注意力权重，显示哪些图像区域和哪些基因特征存在强关联，从而赋予模型一定的可解释性。

病理图像与基因组数据的融合催生了放射基因组学的对应概念——病理基因组学。通过融合，深度学习模型可以从病理图像中预测诸如癌症的分子分型（例如 IDH 基因突变状态），将真实测得的基因

数据与图像一起用于模型训练,可以显著提高预测可靠性。Ding 等[68]开发了多模态 Transformer(PathOmics),将病理图像的区域块嵌入和对应的多基因表达签名一同输入 Transformer 编码器,使模型同时关注图像模式和基因模式,用于结直肠癌患者的生存预测并获得了较传统模型更好的区分度,说明病理和组学的结合能更好地表征肿瘤微环境和分子状态。

除了基因组,病理图像还可与临床信息融合。例如在乳腺癌诊断中,将病理图像特征与患者的激素受体状态、肿瘤大小等临床指标结合,可以训练出比单纯基于图像的模型更准确的预后模型,并且具有一定解释性(如模型会将淋巴细胞浸润程度和患者肿瘤分级等因素结合考虑)[69]。此外,在数字病理领域,一些研究结合了多模态的组织染色结果,如将 H&E 常规染色图像与 IHC 免疫组化标记物的定量结果结合,用多任务学习同时预测组织学特征和分子标志,为辅助诊断提供更全面依据。

总之,数字病理图像的融合拓展了对疾病的多尺度认知——将宏观组织学与分子生物学相连接。挑战在于病理图像的数据量和超高分辨率,使得融合方法需要在效率和效果之间权衡。目前较为成功的策略是先压缩图像信息(通过提取特征或构建图结构)再进行融合。在可预见的未来,随着算力和算法的进步,直接在像素级进行跨模态联动分析(例如结合空间转录组,每个像素或细胞既有图像信息又有基因表达)将成为可能,这将极大推动精准病理诊断的发展。

2.4 临床 EHR 与文本数据

EHR 包含患者诊疗过程中的大量异构数据,包括结构化数据(人口统计信息、检查检验结果、用药记录等)和非结构化数据(医生的文字记录、出院小结等临床文本)(图 6)。EHR 数据反映患者的临床全貌,与影像、组学一起融合可以提供诊断和预后的关键背景信息[70]。例如,两位胸部 CT 影像表现相似的患者可能因病史不同(一个有长期吸烟史,一个无长期吸烟史)而风险有别,这种信息可在 EHR 中找到。因此,EHR 融合在临床决策支持系统(clinical decision support system, CDSS)中扮演重要角色。

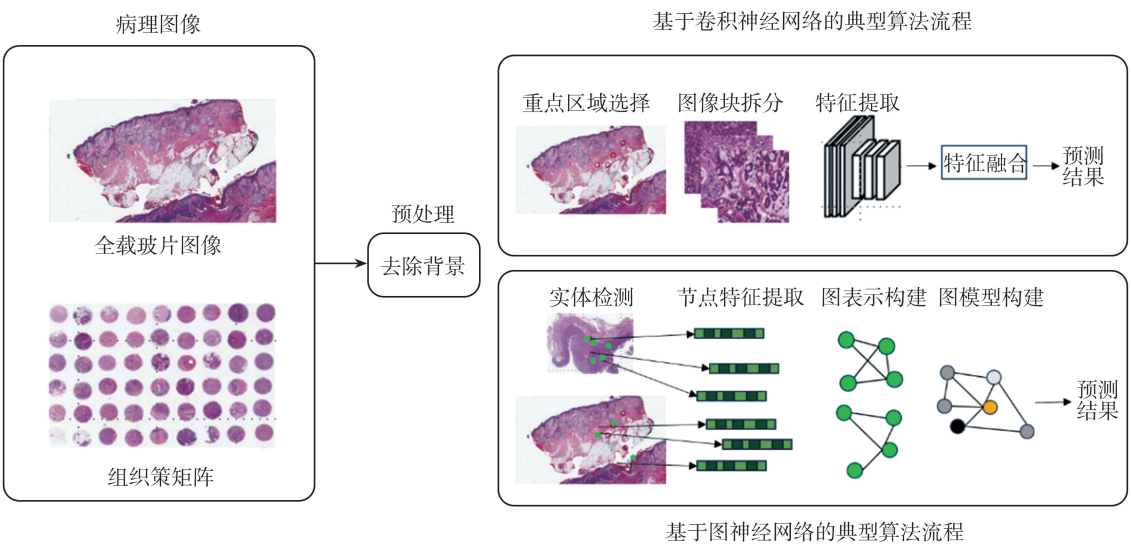

图 6 EHR 数据的模态(结构化与非结构化)
Figure 6 Modalities of EHR data (structured & unstructured)

结构化临床数据(如血压、血糖、化验指标)往往是时间序列数据或多次测量值。融合时需要考虑时间动态和纵向信息。一些研究使用 RNN 或 Transformer 处理这些时间序列数据,再与其他模态结合。例如,Patharkar 等[71]利用 MIMIC-IV 数据库,将 ICU 患者的生理监测时间序列、实验室检验结果与胸片影像、放射科报告三种数据融合,通过深度模型显著提高了 ICU 病死率预测的准确性。

临床文本是 EHR 中信息量极大却最难利用的部分,通常包括医生的自由文本记录(病史、查房记录、手术记录等)。近年来自然语言处理(natural language processing, NLP)的进步使得从临床文本中提取有用特征并与其他数据融合成为可能[72]。一种直接的方法是使用经过临床语料预训练的 BERT 模型,将文本转换为向量嵌入,再与其他模态特征拼接输入模型。有研究提出多模态 Transformer框架,将入院记录文本经过ClinicalBERT

编码为文本嵌入,与结构化 EHR 特征一起输入 Transformer,预测 ICU 患者的院内死亡风险[7]。在 MIMIC 数据集上的验证表明,该融合模型的 AUC 达到 0.877,显著优于只用数值特征的模型。

除了深度学习,还有一些研究利用规则或知识图谱融合 EHR 信息。将 EHR 中的诊断编码和化验结果通过查询医学知识图谱获取关联病症,再与影像特征结合改进诊断。这属于前述图结构方法在临床数据融合上的应用。无论采用何种方式,EHR 融合的一个关键挑战是数据质量与标准化:EHR 数据通常充满噪声、缺失和错误,需要充分的预处理和标准化(如单位换算、术语映射)才能与其他模态数据匹配。另外,EHR 涉及敏感的个人信息,隐私保护也需要考虑。有研究探索在联邦学习框架下融合多个医院的 EHR 与影像数据,即数据不出本地,仅共享模型参数,以保护患者隐私[73]。

2.5 可穿戴设备与移动健康数据

随着移动健康技术的发展,来自可穿戴设备和移动传感器的数据成为医疗数据的新来源。智能手环、可穿戴胸贴、智能手机 App 等可以持续采集心率、血压、血氧、运动步数、睡眠等大量生理和行为数据。这些数据与临床传统数据结合,有助于实现对患者的全程、个性化健康管理。例如,对于心衰患者,融合植入式监测器的压力数据与定期随访的检查结果,可提前发现心功能恶化迹象并及时干预[74]。

可穿戴数据的特点是连续性强、采样频率高,但噪声也大。常见处理方法是提取统计特征或频域特征,再与其他数据融合。用一周的心率数据计算心率变异性等指标,结合患者的遗传风险评分,预测某药物的心脏不良反应风险。近年来也出现将时序信号直接与影像或组学等数据一同输入深度模型的尝试(图 7)。一项针对癫痫发作预测的研究融合了可穿戴设备记录的 EEG 脑电信号和患者脑部 MRI 结构特征,通过 1D 卷积提取 EEG 特征、3D 卷积提取 MRI 特征,然后在融合层结合,显著提高了癫痫发作检测的敏感度[75]。

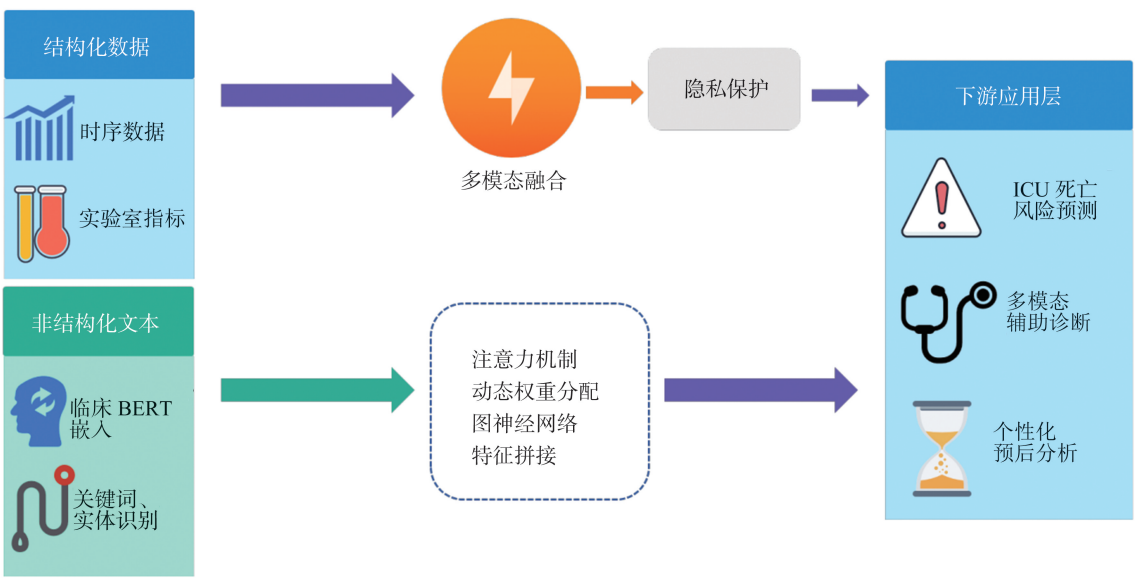

图 7 面向移动设备产生的模态数据

Figure 7 Modal data generated from mobile devices

移动健康数据还包括环境和生活方式数据,如空气质量、地理位置、饮食记录等。这些也可看作一种模态。在哮喘研究中,融合患者可穿戴测得的呼吸频率、GPS 得到的环境污染暴露,以及其遗传易感性评分,用因果图模型分析哮喘发作的触发因素[76]。可穿戴数据与传统医疗数据融合的挑战还在于不同数据源不同步:临床数据通常是不定期的离散事件,而可穿戴数据是连续不断的流,需要开发算法在时间和语义上进行对齐[77]。协同训练和多视角学习可以利用可穿戴数据的自监督信号(如心率与运动的关联)来辅助提升临床模型的稳健性。

随着远程医疗和居家监测的普及,可穿戴数据融合将越来越重要。从长期看,这类数据可以帮助建立个体化的基线,一旦健康指标偏离各自基线即可及时预警和干预。如何在融合模型中有效利用连续的个体纵向数据,同时过滤掉日常噪声,是需要进一步研究的问题。但可以肯定的是,将可穿戴和移

动健康数据纳入多模态融合,有望实现真正的全方位医疗图景,推动医疗模式从院内诊疗向院外健康管理的一体化发展。

3 数据融合策略与训练技术

多模态数据融合的实现途径多种多样。从模型架构上看,按照信息融合发生的阶段,可分为早期融合、中期融合和晚期融合三种基本策略[78]。除此之外,还有一些针对多模态特点设计的训练技术和机制,例如协同训练、模态对齐、联合嵌入以及注意力机制等,能够进一步提升融合效果。

3.1 早期融合(数据层融合)

早期融合指在进入模型学习阶段之前就先将不同模态的数据连接在一起,形成一个综合的输入(图 8)。例如最常见的是特征级拼接,即直接把各模态的原始特征向量或提取的表示向量串联成一个长向量输入模型。在深度学习框架中,早期融合通常意味着模型的第一层就同时接收所有模态的数据。例如,一个多模态神经网络的输入层接受拼接后的多模态特征,然后通过共享的后续网络层进行预测[79]。这等效于假设存在一个统一的特征空间,其涵盖所有模态。

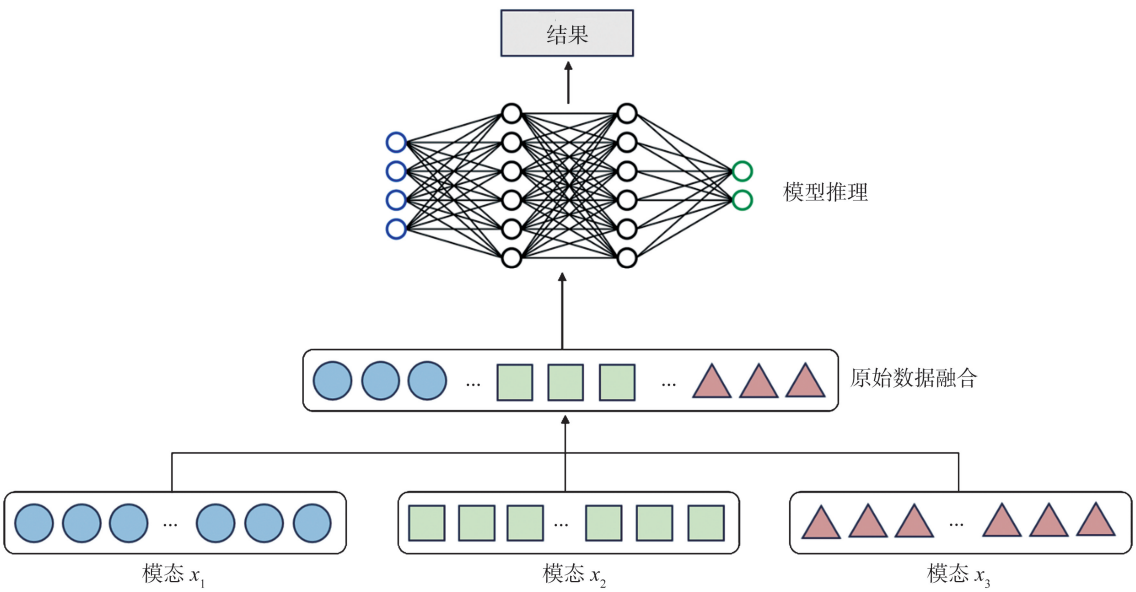

图 8 早期融合
Figure 8 Early fusion

早期融合的主要优点是实现简单且融合彻底。由于所有模态数据在模型初始即已结合,模型可以在很低层次就开始学习模态之间的关系,这有利于跨模态关系主要体现在低级特征上的任务。在将心电信号和血压波形一起用于分类某种疾病时,早期融合允许模型在卷积层就同时“看到”两种波形,可能检测出同步变化等模式。而且早期融合避免了需要为每个模态单独设计子模型,降低了架构设计难度[80]。

然而,早期融合也存在明显问题。① 如果不同模态的数据分布尺度差异大或冗余特征多,简单拼接可能使模型训练困难,需要预先处理(如归一化或降维),以避免某模态主导训练。② 早期融合缺乏灵活性,模型必须在统一空间同时适配所有模态特征,可能无法捕获那些需要更高层抽象才能显现的跨模态关联。基因突变对影像表型的影响也许只有经过复杂网络才能体现,如果在低层就简单拼接,模型早期可能更关注模态内模式,难以发现跨模态的高阶关系[81]。③不同模态采样频率不同(如可穿戴设备连续数据 *vs.* 医院记录离散数据),早期融合直接拼接时间序列和单时间点数据会不协调[82]。④ 早期融合对缺失模态较为敏感。如果某模态数据缺失,拼接向量就不完整,需要额外处理(如填充缺失值或采用多入口网络等)。

尽管如此,早期融合在一定条件下可以取得与更复杂融合策略相当的效果。尤其在模态之间强相关且噪声较低的情况下,早期融合的简单策略往往足够。在放射影像的多序列(如 MRI 的 T1 加权、T2 加权等)联合诊断中,不同序列本质上是同一解剖结构的不同成像对比,直接将多序列图像的通道维度叠加输入 3D-CNN 往往效果良好[83]。因此,在实际应用中,可以先尝试将早期融

合作为基线方法,然后根据需要再考虑更复杂的融合架构。

3.2 中期融合(特征层融合)

中期融合也称特征层融合或联合融合,是指各模态数据先分别经过各自的子网络提取中间表示,然后在中间层将这些表示融合,再继续经过若干共享网络层,得到最终输出[84](图 9)。相比早期融合的"一股脑混合",中期融合允许模型先学习单模态内部特征,再学习跨模态特征。这类策略应用非常广泛,因为它提供了在融合时机和方式上的灵活性。

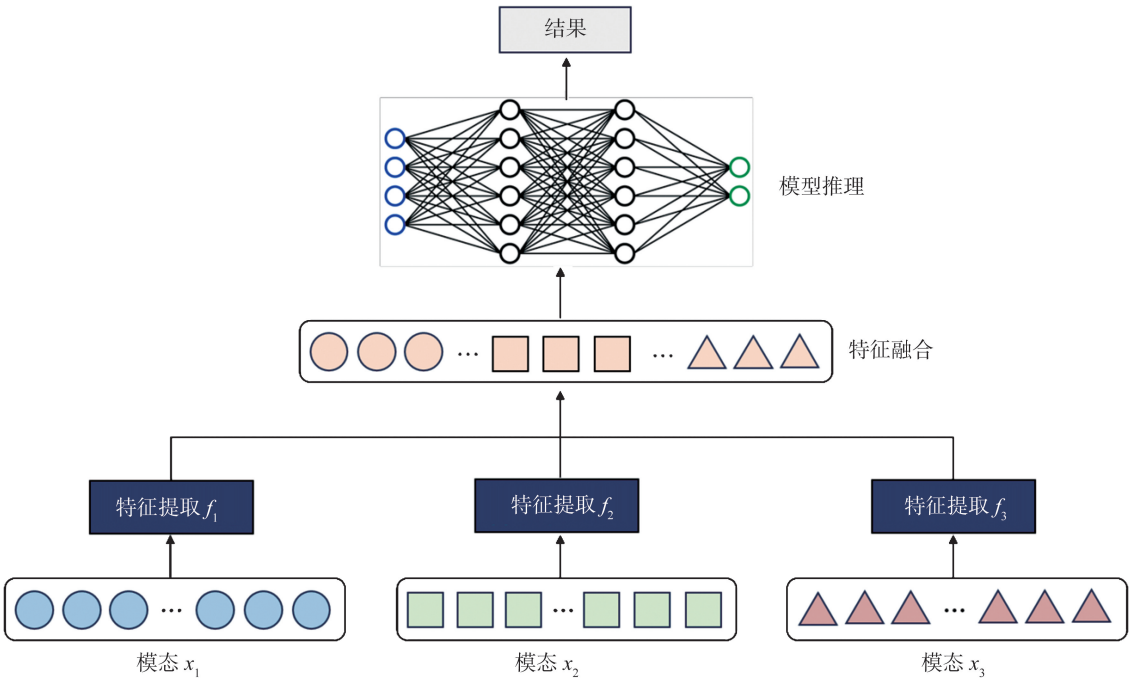

图 9 中期融合

Figure 9 Middle fusion

根据融合发生的深度和方式,中期融合可以细分为多种模式:如果各模态提取出特征后直接拼接送入分类器,这有时被称为特征晚期融合,但本质上仍属于中期融合的范畴。而联合中期融合则指在拼接特征后还经过共享网络层进一步学习联合表示。此外,还有渐进融合的概念,指设置多个融合节点,逐步将模态加入融合序列,例如先融合 A 和 B 模态,经过几层网络后再融合 C 模态。这种方式在模态较多且两两相关性不一致时很有用:高度相关的模态可以早点融合,而弱相关的模态稍晚融合。

中期融合的优势在于灵活且高效。通过为每个模态配备适合其数据类型的子网络(如图像用 CNN、文本用 Transformer、表格数据用全连接网络),可以充分提取每种模态的边际特征,然后再在高层实现模态间的信息交换。这不仅更贴近模态间关系的自然层次(许多跨模态关联在高层语义上才出现),也能缓解异质模态直接融合时的冲突[85]。实证研究表明,在许多任务上,中期融合比早期融合表现更佳。在疾病诊断中,先让影像特征经过几层 CNN 抽象后再与遗传数据融合,能够捕获影像的病灶级特征再结合遗传风险,从而效果优于直接融合原始影像像素和基因数据[86]。中期融合还方便处理部分模态缺失:如果某模态缺失,可以用其子网络输出默认值或直接忽略该分支,不影响其他模态子网的提取和后续融合。

需要注意的是,中期融合也有挑战。① 架构设计复杂度,即需要决定每个模态子网的深度和结构,以及在哪一层进行融合,这通常依赖经验和实验调整。② 中期融合模型参数量往往较大(因为有多套子网络),在训练数据有限时可能过拟合,需要采用正则化或部分权重共享等方式改进。③ 中期融合虽然理论上更强大,但优势在某些任务上未必显著。对于一些简单任务,早期融合或晚期融合有时也能取得相近性能,而模型更简单。因此应根据任务复杂度权衡架构设计,不必一味追求复杂。

总体而言,中期融合因其在融合时机和方式上的灵活性,已成为当前多模态深度学习的主流选择。特别是联合中期融合(各模态特征拼接后还有共享网络精炼)的方式,能充分学习模态交互,被广泛应用于医学多模态任务中。多数先进方法倾向于在中期融合采用联合表示学习,有效建模不同生物层次间的复杂相互作用[8]。因此,在设计多模态模型时,中期融合通常是一个合理的起点,再根据数据和任务特点做定制化调整。

3.3 晚期融合(决策层融合)

晚期融合也称决策级融合,其特点是各模态分别训练子模型(各自输出预测或决策),最后在决策层将这些输出合并(图 10)。例如,对于图像和文本两种模态,可以各自训练一个分类器输出疾病概率,然后取平均或投票作为最终结果。晚期融合本质上类似集成学习,把不同模态的模型当作“专家”,再通过某种规则综合判断[8]。

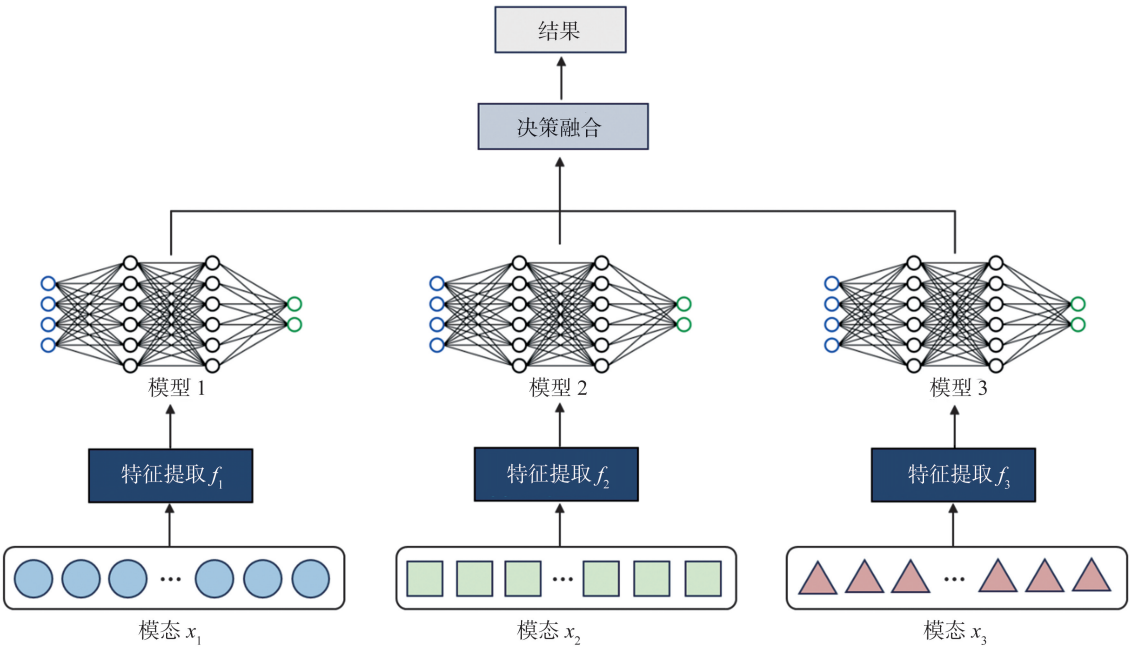

图 10 晚期融合

Figure 10 Late fusion

晚期融合的一个显著优点是模块化和独立优化。每个模态子模型都可以针对本模态进行最优训练和调参,不需要在训练过程中考虑其他模态的数据分布。这种划分降低了模型的耦合复杂度,也减少了多模态数据同步的需求。在实际应用中,当不同模态的数据具有不同来源或采集时间,晚期融合可以方便地整合数据,例如将影像分析模型的结果与化验指标评分相加得到最终分数。由于各子模型相对独立,可扩展性也好,增加新模态只需训练一个对应的新模型再加入融合,不必重新训练整个系统。

此外,晚期融合允许异构模型并存,子模型可以是不同类型的模型(如一个随机森林结合一个神经网络),不影响最终融合实现,这对于合并已有的成熟单模态模型非常有用。在医学 AI 中,一些经典的单模态模型(比如 CHA_2DS_2-VASc 评分这类基于规则的风险评分模型)可能已经过充分验证,而新模态可能采用深度学习模型,那么可以在决策层以加权方式融合规则评分和深度模型输出。这种做法在提高性能的同时也兼顾了一定的临床可接受性,因为融合结果的一部分来自医务人员熟悉的传统模型。

晚期融合的另一个优点是鲁棒性。如果某个模态数据缺失或质量差,仍可用其他模态的子模型输出给出结果,不至于使整个系统失效。各子模型错误往往不相关,通过融合平均可以降低总体误差。这类似 bagging 思想,在一定程度上提升了泛化性能。不同模态模型由于输入差异,错误模式往往不同,将它们的预测结果进行平均或投票,可以提高预测的稳定性。

当然,晚期融合的缺点也很明显。由于在最后一步才融合,各模态间的交互关系无法在模型内部直接学习。模型只是简单混合最终判断,无法捕获模态 A 的某种模式与模态 B 的某种模式组合产生的效应。因此在存在显著交互作用的任务上,晚期融合可能表现不佳。此外,如何设计最优的决策融合规则也是问题。简单平均可能不是最优,常用的方法包括加权平均(根据验证集性能给每个模态模型赋权)或训练一个元模型做裁决。例如,可以训练一个逻辑回归模型,以各子模型的预测值作为输入,学习最佳加权。这种方法提高了融合灵活性,但需要额外的训练数据。还有一种思路是利用置信度,如果某模态模型对某样本预测不确定性很高,可以降低其权重,从而自适应地融合,但实现可靠的置信度估计并不容易。

晚期融合经常作为 baseline 或提升性能的集成

手段出现。在一些比赛或挑战中，参赛队伍会训练多个模态的模型，然后做晚期融合作为最终提交，以提高成绩的稳定性。这说明晚期融合在实际系统的实用价值。然而，如果目标是深入挖掘模态间的交互关系，晚期融合显然不如中期融合。因此学术研究更多关注中期融合，但在产品落地中，为了利用已有模型或方便模块化更新，常采用晚期融合。例如医院可能已有独立的心电图 AI 和影像 AI，直接用晚期融合将两者输出结合即可，无需重写一个“大一统”的模型。

3.4 模态对齐与联合嵌入

模态对齐和联合嵌入是为了解决不同模态异质性所采取的关键策略。所谓模态异质性，指不同数据模态在表现形式、取值空间上截然不同，直接融合容易“风马牛不相及”。因此，需要将它们投射到某种共同空间或找到它们之间的对应关系[87]。

模态对齐可以分为显式对齐和隐式对齐[79]（图 11）。显式对齐是指在数据层面明确建立模态间的一一对应关系。在影像-文本融合中，显式对齐就是知道某段报告文本对应某张医学影像[29]；在多组学数据中，显式对齐意味着知道哪些基因表达测定对应哪个影像区域或细胞样本。有了这种配对关系，融合模型就可以利用监督信号，例如用对比学习拉近正确配对的模态表示距离。隐式对齐则是不直接给定对应，而是通过模型学习去自动发现。例如，使用注意力机制让模型自己去寻找影像区域与文本描述句子之间的对应关系[88]。隐式对齐往往结合联合嵌入来实现。

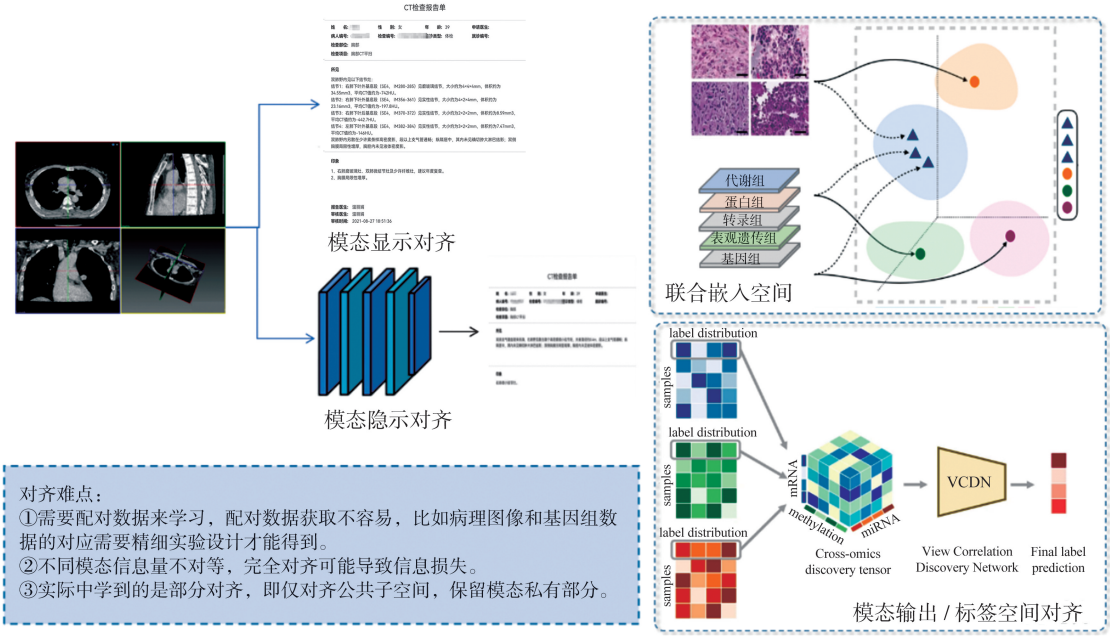

图 11 模态对齐

Figure 11 Modality alignment

联合嵌入空间是指学习一个向量空间，使得投影到该空间中的不同模态的数据可以直接进行比较和运算[29]。最典型的例子是 OpenAI 的 CLIP 模型在自然图文领域的成功，其图像和文本被映射到同一个特征空间，通过对比训练使得描述同一图像的文本和图像特征距离很近，不对应的图文距离较远。医疗领域也有类似工作，如前述 ConVIRT 将放射影像和报告嵌入到公共空间，后续即可通过简单的向量检索实现“以文找图”或“以图找文”[30]。在联合嵌入空间中，不同模态的表示具有可比性，实现了对齐，即相同语义的内容会聚在一起、不同语义的内容分散开来。

构造联合嵌入通常需要对比损失或重构损失的辅助。前者如上所述，通过最大化正确配对样本表示间的一致性来实现[30]；后者则如多模态自编码，训练一个模型同时能够从嵌入重建出模态 A 和模态 B 的数据。为了避免仅记住单模态细节，需要引入正则确保嵌入进行了压缩或两个模态的重建相互约束。在语言-时间序列数据融合中，可训练一个双头自编码器，使得共有的潜在表示既能重建时间序列，也能生成相应的文本报告，迫使表示包含两种模态的公共信息。

模态对齐还可以发生在输出/标签空间上对齐。MOGONET 中使用的视图关联发现网络（view correlation discovery network，VCDN）就是在标签空间对齐：先得到每个模态对于各类别的预测分布，然后

寻找让不同模态预测之间关系最匹配真实标签的融合结果[5]。这种做法对于分类任务有效,可理解为在输出端建立模态间的一致性。模态对齐的好处在于可以利用跨模态的一致性来训练模型,一张 X 光片和其对应的文字报告本质上传递的是同一信息,只是形式不同。如果模型能学会对齐,那么即使报告没有人工标签,也可作为 X 光的弱标签,从而利用额外的未标注数据,前文提及的自监督对比学习正是利用了这种原理[29]。此外,良好的对齐能让模型具备多模态推理能力。比如给定一种模态的数据,可以在联合空间中找到最近邻的另一模态实例,从而实现跨模态检索、跨模态诊断支持等功能。在临床实践中,这意味着医生可以输入一段症状描述,系统从影像数据库中找出相似病例的影像并提示可能的诊断。

当然,实现模态对齐也有难点。首先,需要配对数据来学习,有时获取完美配对的数据并不容易,比如病理图像和基因组数据的对应需要精细的实验设计才能得到。其次,不同模态的信息量不对等,完全对齐可能导致信息损失,例如影像包含丰富细节,而文本可能只描述异常,如果要求二者完全对齐,则影像细节可能无法全部表达。因此实际中往往学到的是部分对齐,即仅对齐公共子空间,同时保留模态特有部分。这与上文提到的共性-特异性分解思想是一致的。可将二者结合,首先对模态特征做共特异分解,只对齐其中的共性部分,从而在保证对齐的同时也保留模态特有优势。

3.5 协同训练与多视角学习

在多模态任务中,不同模态往往可以互相监督彼此,从而缓解标注数据不足的问题。协同训练是一种半监督学习策略,它假设每种模态(或特征子集)可以单独训练一个分类器,并且各模态分类器的错误是条件独立的。通过让不同模态的模型互相教导对方未标注样本的高置信度预测,可以扩大训练集,提高整体性能。在医学场景,如果某模态标注困难,可以用另一模态较容易获得的标签来训练初步模型,再互相迭代提升。比如,用影像模型为未确诊患者预测疾病风险,挑选高置信度的部分作为伪标签补充给基因数据模型训练,反之亦然,从而使两个模态的模型精度共同提高。

多视角学习与此类似,将不同模态视为同一任务的不同视角。模型可设计为鼓励不同视角的一致性。例如,在多模态学习中加入视角一致损失,让模型的各模态子网络在对同一样本输出时保持某种相似性或相关性。具体方法包括最大化不同模态表示之间的互信息,或最小化它们之间的距离。当视角之间存在内在联系时,这种约束有助于模型学习到模态间的共享本质。

需要强调的是,协同训练和多视角学习要求模态之间存在一定冗余或互补,这样一个模态的信息才能帮助提升另一个模态。否则,如果模态完全独立,强行协同可能引入噪声。实践中常结合验证集性能来判断协同的收益,并注意避免模型之间的相互依赖导致“信息泄露”(如一个模型过度拟合另一个模型的输出)。

4 多模态融合实例分析

多模态医学数据融合的最终目的在于提高临床决策和科研发现的质量。近 5 年来,多模态融合技术在多个重要的医学应用场景中取得突破性进展。下文围绕几个具有代表性的场景,介绍融合方法如何在实际问题中发挥作用,并给出相应的研究案例。

4.1 癌症预后与生存分析

癌症预后评估需要综合考虑肿瘤的分子特征和患者的临床情况,是天然的多模态问题。传统预后模型往往仅依赖临床分期等单一信息,而融合多组学和影像数据可以获得更精确的生存预测和复发风险评估[89](图 12)。

多模态融合的一个突出方向是将病理组织学+分子数据的融合用于预后模型。Chen 等[90]开发了 pathomic fusion 模型,将病理图像、基因组和临床变量在中期融合,通过联合注意力机制学习一个三模态共享表示,用于预测肺癌患者生存。除了病理影像,放射影像+组学(即放射基因组学)在预后中的应用也很活跃。将 MRI 的肿瘤纹理特征与基因表达亚型融合,可以更准确地预测胶质瘤患者的生存时间[91]。Magbanua 等[92]将 CT 影像的影像组学特征与血液中的循环肿瘤 DNA 突变负荷结合,提高了对非小细胞肺癌化疗效果的预测,为预后评估提供了新的生物标志组合。

融合模型不仅提供预测结果,还能挖掘潜在的预后因子。例如,有研究使用多模态图注意力网络预测乳腺癌生存时发现,某些影像区域(如肿瘤边缘)和基因标志(如促细胞增殖基因)共同构成高风险特征[47]。这些发现与已知的肿瘤边缘浸润现象和增殖驱动机制相吻合,增强了模型的可信度。

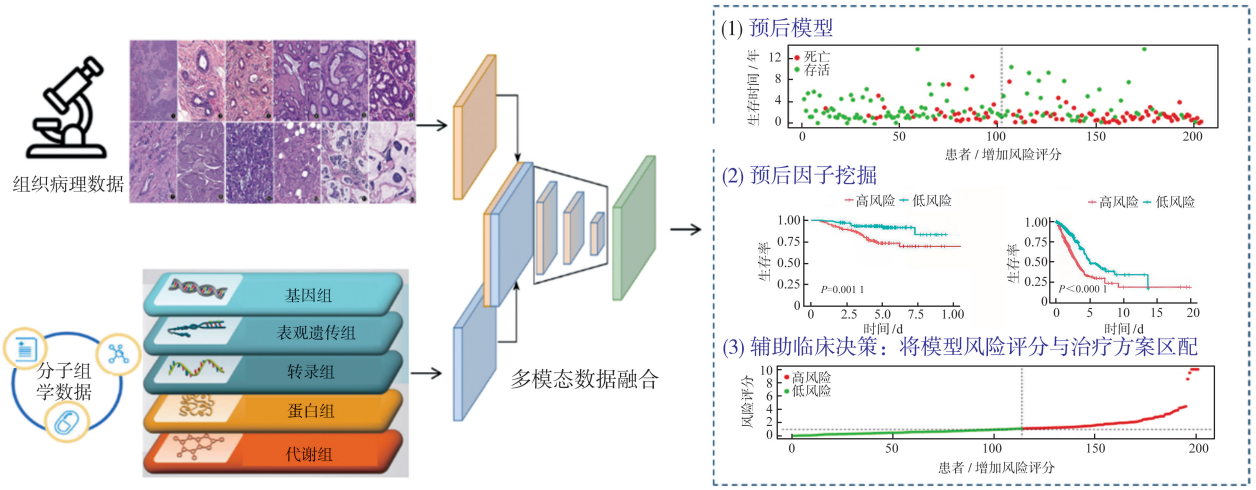

图 12 基于多模态融合的预后分析

Figure 12 Prognostic analysis based on multimodal fusion

值得一提的是,多模态预后模型也可辅助治疗决策。通过将模型风险评分与治疗方案相匹配,可以筛选出需要更积极治疗的患者亚组。例如,如果融合模型预测某患者术后有高复发风险,则医生可考虑加强辅助化疗。在乳腺癌等需决策辅助治疗的疾病中,这种精细分层尤为重要。近年有研究将 21 基因表达风险评分与传统临床指标和病理图像融合,训练模型建议是否需要化疗,其性能超过单用 21 基因或单用临床分期的决策[1]。这表明多模态融合有潜力直接影响治疗方案的制定,实现更个性化的干预。

总体而言,在癌症预后场景下,多模态融合带来的精度提升和洞察加强已经得到初步验证,挑战在于如何获得大规模的配对多模态数据(如 TCGA 提供了多类型数据但整体样本量有限)。未来,更多前瞻性研究和真实世界数据的整合将推动这些融合模型真正进入临床实践。通过多中心合作收集样本,实现病理、影像、组学和随访数据的联动分析。一旦融合模型充分验证其可靠性,有望作为新一代预后评分系统应用于临床,为每位患者制定精准的治疗和随访方案。

4.2 疾病诊断分型与亚组发现

除了预后外,多模态融合在疾病的诊断分型和未知亚组发现上也展现出巨大优势。这里的分型指根据多源信息对疾病进行更细致的分类(如分子分型、亚群划分),以指导精准治疗(图 13)。

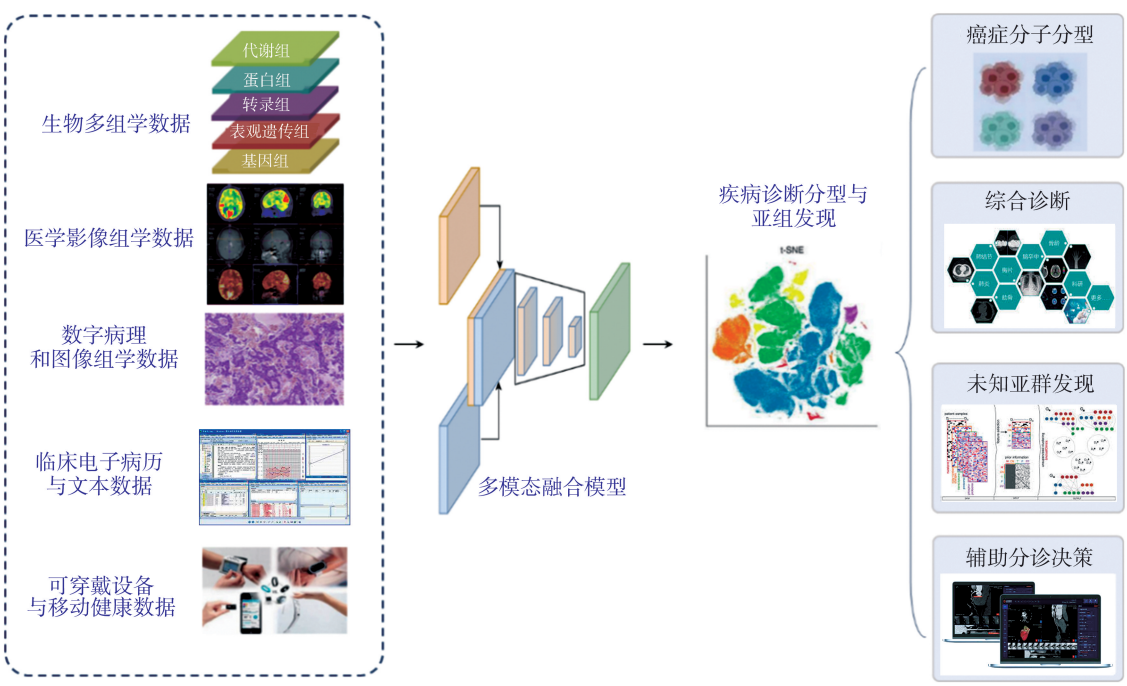

图 13 基于多模态融合的疾病诊断分型

Figure 13 Disease diagnosis and subtyping based on multimodal fusion

癌症分子分型是精准医疗的典型例子。过去分型主要依靠组织学,现在可以综合影像和基因信息。以胶质瘤为例,其分型依赖 IDH 突变和 1p/19q 共缺失等分子标志,但这些信息需要分子检测才能获得。研究者开发了深度融合模型,输入 MRI 影像和患者年龄等临床数据,输出预测的 IDH 突变状态和 1p/19q 状态[93]。结果显示,多模态模型预测胶质瘤分型的准确率达到 90%以上,接近实际分子检测,从而有望通过无创手段获知分型信息。类似地,也有将放射影像与临床数据融合预测肝癌、肺癌等肿瘤分子亚型的工作,以辅助筛查和治疗选择[94]。

在复杂疾病综合诊断方面,多模态融合有助于提高疑难病例的诊断率。例如系统性红斑狼疮(systemic lupus erythematosus, SLE)的确诊需要综合临床症状、实验室自身抗体结果、病理检查等多方面证据,单一 AI 模型难以处理如此多元的信息。科研人员尝试构建融合系统,输入患者的症状文本、免疫指标(如 ANA 抗体滴度)、肾活检病理图像等,由多模态网络综合判断是否为 SLE 及其活动程度。结果显示,其诊断准确率比风湿科平均水平更高且稳定。目前此类应用多在研究阶段,但前景诱人——AI 可辅助诊断复杂综合征,降低漏诊误诊。

发现未知亚群也是多模态融合的重要应用。通过无监督或弱监督的机器学习,可以在多模态数据中发现新的疾病亚型。Suter 等[36]通过融合多种组学和蛋白数据发现了肝癌的三种新分子亚型,每种亚型在生物通路激活和预后上都有差异。这些亚型部分印证了已有研究,但也提供了新见解,如某亚型表现出特殊的磷酸蛋白网络激活,提示了新的潜在药物靶点。多模态融合能发现单一模态无法揭示的潜在分类,因为不同模态提供了额外的区分维度——仅靠基因数据可能看不出两群体差异,但结合代谢数据可能将它们明显区分开。

辅助急诊分诊决策也是一个应用点。例如急性胸痛患者通常会进行心电图、抽血化验、胸部 CT 等多项检查。融合模型可以综合这些数据,快速判断可能的病因(如心肌梗死、肺栓塞、主动脉夹层等),给出分诊建议(如立即心导管检查或安排 CT 肺动脉造影),从而节省宝贵的抢救时间。

综上,多模态融合已经在准确分型和复杂诊断方面展现出显著优势。在医疗实践中,这意味着患者可以少走弯路,接受更有针对性的治疗。例如乳腺癌患者在手术前就通过融合模型预测分子分型,可同步制定相应的内科治疗方案;多学科团队对疑难疾病进行讨论时,由 AI 提供综合分析意见,可减少争议和遗漏。当然,模型在这些场景下需要达到极高的可靠性,并且其建议往往只是辅助,最后决策仍由医生做出。但从技术角度,多模态融合正在逐步突破以往各自为战的诊断流程,使之朝全面集成的方向前进。

4.3 病理图像辅助分析

在数字病理领域,多模态融合技术除了用于预后和分型外,还有一些特殊的应用场景,例如肿瘤分级和治疗反应预测等。这里重点讨论病理图像结合其他数据进行辅助分析所发挥的作用(图 14)。

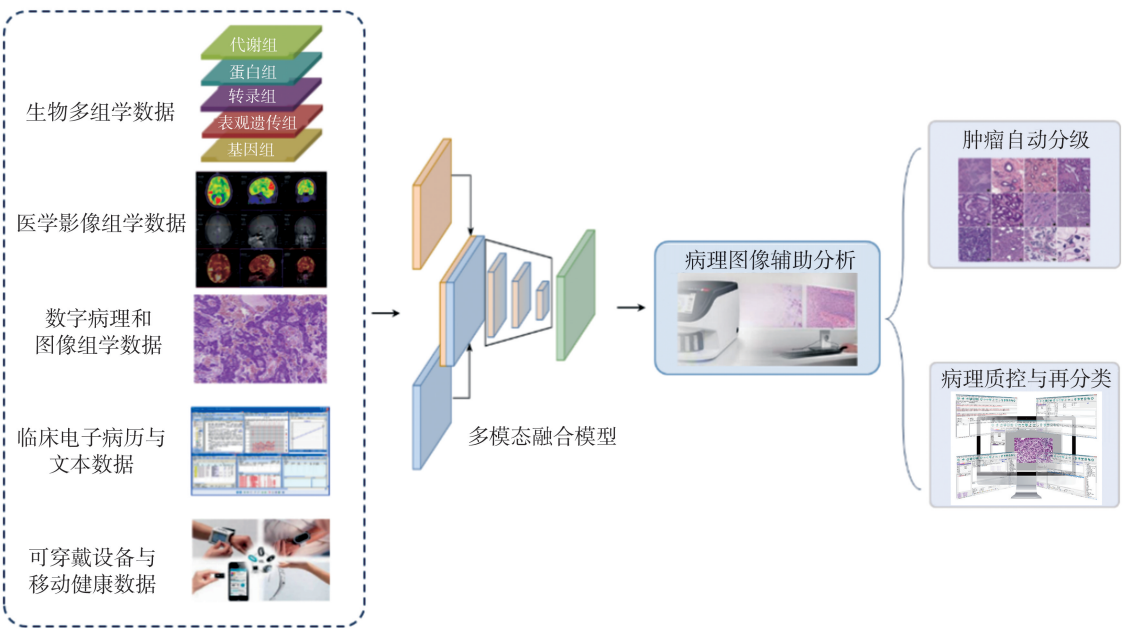

图 14 基于多模态融合的病理图像辅助分析

Figure 14 Pathological image analysis assisted by multimodal fusion

4.3.1 肿瘤自动分级

病理学上，一些肿瘤需要根据细胞异型性、分裂象等特征进行分级，这往往影响治疗选择，然而单纯依靠图像定量这些特征有时不够稳定。如果引入分子数据和临床信息，可以提高分级的准确性和客观性。比如针对前列腺癌的 Gleason 评分，单看切片在临界案例上主观差异较大。一项研究融合了病理图像的深度特征和患者血清 PSA 水平，通过联合模型给出 Gleason 等级预测，结果与资深病理医师诊断高度一致，并减少了中间级别的不确定性[95]。这说明融合临床指标（PSA 在一定程度上反映肿瘤负荷）能帮助图像模型更好地区分 Gleason 评分中的临界级别。另外，乳腺癌细胞核异型性的评分也尝试结合基因表达亚型信息，使模型更倾向于将 Basal-like分子亚型的病例判为高级别，因为该亚型通常对应三阴性乳腺癌，生物学侵袭性高。

4.3.2 病理质控与再分类

实际诊断中，初始病理诊断并非总是准确的，常需在获得更多信息后修正。AI 可通过融合模型对初诊结果提出质疑并辅助质控。例如，弥漫大 B 细胞淋巴瘤和伯基特淋巴瘤在形态上类似，但基因特征不同。融合病理图像和基因表达的模型可以对形态学诊断进行二次检查，若发现不一致则警示病理医生进行复核。这个过程类似于给疑难或易混淆疾病加上双保险。当前已有基于影像+基因融合的模型能够区分几种形态近似的脑瘤并提示可能的分类错误[96]。

4.3.3 诊断升级与分子预测

病理图像作为终极诊断证据，与多模态融合结合产生了强大的功能。一方面，它提供了组织学层面的真相，使得融合分析更可信、更贴近临床；另一方面，其他模态赋予病理图像更多维度的信息解读，使传统显微镜难以察觉的分子差异变得可见。可以预见，未来的病理科报告中可能不仅有组织学描述，还会附上一段由 AI 根据多模态数据推断出的分子特征和预后预测。例如："本病例肺癌组织学诊断为腺癌；AI 推断其 EGFR 基因突变概率 90%，肿瘤浸润淋巴细胞水平高，提示对免疫治疗的应答可能性大。"这种报告将极大提升病理科在精准医疗中的核心决策支持作用。

5 结论与展望

多模态医学数据融合技术在过去 5 年中取得了令人瞩目的进展。从方法论上，深度学习模型（如 Transformer、GNN、自监督学习）展现出强大的跨模态表示学习能力，统计模型（如贝叶斯融合、潜变量分解）提供了稳健的理论支撑，图模型则将先验知识和结构信息引入融合，实现了"数据+知识"的结合。在数据层面，测序、成像、传感等技术的发展使得同一患者的多组学、医学影像、临床记录乃至日常行为数据能够被同时获取并关联分析，为全景式诊疗铺平了道路。各类融合策略（早期、中期、晚期融合）与注意力机制、协同训练等技巧的灵活运用，让模型可以针对具体任务调整融合方式，在精度和可解释性之间取得平衡。更重要的是，这些技术在多个关键应用场景中展现了临床价值，提高了癌症预后预测的准确度，发现新的疾病分子亚型，辅助疑难病例的诊断决策，以及实现实时监护预警和个性化治疗方案制定等。多模态融合正在推动医疗模式从经验驱动转向数据驱动、从群体统计转向个体精准化。

参考文献：

[1] Teoh JR, Dong J, Zuo XW, et al. Advancing healthcare through multimodal data fusion: a comprehensive review of techniques and applications[J]. Peer J Comput Sci, 2024, 10: e2298. doi:10.7717/peerj-cs.2298

[2] Krones F, Marikkar U, Parsons G, et al. Review of multimodal machine learning approaches in healthcare[J]. Inf Fusion, 2025, 114: 102690. doi:10.1016/j.inffus.2024.102690

[3] Kumar S, Rani S, Sharma S, et al. Multimodality fusion aspects of medical diagnosis: a comprehensive review[J]. Bioengineering, 2024, 11(12): 1233. doi: 10.3390/bioengineering11121233

[4] Chaabene S, Boudaya A, Bouaziz B, et al. An overview of methods and techniques in multimodal data fusion with application to healthcare[J]. Int J Data Sci Anal, 2025. doi:10.1007/s41060-025-00715-0

[5] Wang T, Shao W, Huang Z, et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification[J]. Nat Commun, 2021, 12(1): 3445. doi: 10.1038/s41467-021-23774-w

[6] Shaik T, Tao XH, Li L, et al. A survey of multimodal information fusion for smart healthcare: mapping the journey from data to wisdom[J]. Inf Fusion, 2024, 102: 102040. doi:10.1016/j.inffus.2023.102040

[7] Lyu W, Dong X, Wong R, et al. A multimodal transformer: fusing clinical notes with structured EHR data for interpretable in-hospital mortality prediction[J]. AMIA

Annu Symp Proc, 2023, 2022: 719-728.
[8] Stahlschmidt SR, Ulfenborg B, Synnergren J. Multimodal deep learning for biomedical data fusion: a review[J]. Brief Bioinform, 2022, 23(2): bbab569. doi:10.1093/bib/bbab569
[9] Zheng Y, Conrad RD, Green EJ, et al. Graph attention-based fusion of pathology images and gene expression for prediction of cancer survival[J]. IEEE Trans Med Imaging, 2024, 43(9): 3085-3097.
[10] Lahat D, Adali T, Jutten C. Multimodal data fusion: an overview of methods, challenges, and prospects[J]. Proc IEEE, 2015, 103(9): 1449-1477.
[11] Krones F, Marikkar U, Parsons G, et al. Review of multimodal machine learning approaches in healthcare [J]. Inf Fusion, 2025, 114: 102690. doi:10.1016/j.inffus.2024.102690
[12] Han X, Chen S, Fu Z, et al. Multimodal fusion and vision-language models: a survey for robot vision[EB/OL]. (2025-04-03)[2025-04-26]. https://arxiv.org/abs/2504.02477v2
[13] 张虎成, 李雷孝, 刘东江. 多模态数据融合研究综述[J]. 计算机科学与探索, 2024, 18(10): 2501-2520.
ZHANG Hucheng, LI Leixiao, LIU Dongjiang. Survey of multimodal data fusion research[J]. Journal of Frontiers of Computer Science and Technology, 2024, 18(10): 2501-2520.
[14] 潘梦竹, 李千目, 邱天. 深度多模态表示学习的研究综述[J]. 计算机工程与应用, 2023, 59(2): 48-64.
PAN Mengzhu, LI Qianmu, QIU Tian. Survey of research on deep multimodal representation learning[J]. Computer Engineering and Applications, 2023, 59(2): 48-64.
[15] 任泽裕, 王振超, 柯尊旺, 等. 多模态数据融合综述[J]. 计算机工程与应用, 2021, 57(18): 49-64.
REN Zeyu, WANG Zhenchao, KE Zunwang, et al. Survey of multimodal data fusion[J]. Computer Engineering and Applications, 2021, 57(18): 49-64.
[16] Rajendran S, Pan W, Sabuncu MR, et al. Learning across diverse biomedical data modalities and cohorts: challenges and opportunities for innovation[J]. Patterns (N Y), 2024, 5(2): 100913. doi: 10.1016/j.patter.2023.100913
[17] Wang T, Li F, Zhu L, et al. Cross-modal retrieval: a systematic review of methods and future directions[EB/OL]. (2025-04-17)[2025-04-26]. https://ieeexplore.ieee.org/abstract/document/10843094/
[18] Sarraf A, Azhdari M, and Sarraf S. A comprehensive review of deep learning architectures for computer vision applications[J]. ASRJETS, 2021, 77(1): 1-29.
[19] Garg M, Ghosh D, Pradhan PM. Multiscaled multi-head attention-based video transformer network for hand gesture recognition[J]. IEEE Signal Process Lett, 2023, 30: 80-84. doi:10.1109/LSP.2023.3241857
[20] Kumar S, Sharma S, Megra KT. Transformer enabled multi-modal medical diagnosis for tuberculosis classification[J]. J Big Data, 2025, 12(1): 5. doi:10.1186/s40537-024-01054-w
[21] Zhou HY, Yu Y, Wang C, et al. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics[J]. Nat Biomed Eng, 2023, 7(6): 743-755.
[22] Nguyen HH, Blaschko MB, Saarakkala S. Clinically-inspired multi-agent transformers for disease trajectory forecasting from multimodal data[J]. IEEE Trans Med Imaging, 2024, 43(1): 529-541.
[23] Khader F, Kather JN, Müller-Franzes G, et al. Medical transformer for multimodal survival prediction in intensive care: integration of imaging and non-imaging data [J]. Sci Rep, 2023, 13(1): 10666. doi: 10.1038/s41598-023-37835-1
[24] Valous NA, Popp F, Zörnig I, et al. Graph machine learning for integrated multi-omics analysis[J]. Br J Cancer, 2024, 131(2): 205-211.
[25] Guo D, Shao Y, Cui Y, et al. Graph attention tracking [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 9543-9552. http://openaccess.thecvf.com/content/CVPR2021/html/Guo_Graph_Attention_Tracking_CVPR_2021_paper.html
[26] Zheng Y, Gindra RH, Green EJ, et al. A graph-transformer for whole slide image classification[J]. IEEE Trans Med Imaging, 2022, 41(11): 3003-3015.
[27] Zheng Y, Conrad RF, Green EJ, et al. Graph attention-based fusion of pathology images and gene expression for prediction of cancer survival[J]. IEEE Trans Med Imaging, 2024, 43(9): 3085-3097.
[28] Huang SC, Pareek A, Jensen M, et al. Self-supervised learning for medical image classification: a systematic review and implementation guidelines[J]. NPJ Digit Med, 2023, 6(1): 74. doi:10.1038/s41746-023-00811-0
[29] Zhang Y, Jiang H, Miura Y, et al. Contrastive learning of medical visual representations from paired images and text [J]. PMLR, 2022: 2-25.
[30] Wang Z, Wu Z, Agarwal D, et al. MedCLIP: contrastive learning from unpaired medical images and text[C]. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2022. https://pmc.ncbi.nlm.nih.gov/articles/PMC11323634/
[31] Taleb A, Lippert C, Klein T, et al. Multimodal self-supervised learning for medical image analysis. In Feragen A, Sommer S, Schnabel J (Ed.) Information

Processing in Medical Imaging. Cham: Springer, 2021: 661-673.

[32] Zong Y, Aodha OM, Hospedales TM. Self-supervised multimodal learning: a survey[J]. IEEE Trans Pattern Anal Mach Intell, 2025, 47(7): 5299-5318.

[33] Ghassemi M, Pimentel M, Naumann T, et al. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data[J]. Proc AAAI Conf Artif Intell, 2015: 446-453.

[34] AlSaad R, Abd-Alrazaq A, Boughorbel S, et al. Multimodal large language models in health care: applications, challenges, and future outlook[J]. J Med Internet Res, 2024, 26: e59505. doi:10.2196/59505

[35] Gygi JP, Konstorum A, Pawar S, et al. A supervised Bayesian factor model for the identification of multi-omics signatures[J]. Bioinformatics, 2024, 40(5): btae202. doi:10.1093/bioinformatics/btae202

[36] Suter P, Dazert E, Kuipers J, et al. Multi-omics subtyping of hepatocellular carcinoma patients using a Bayesian network mixture model[J]. PLoS Comput Biol, 2022, 18(9): e1009767. doi:10.1371/journal.pcbi.1009767

[37] Samorodnitsky S, Wendt CH, Lock EF. Bayesian simultaneous factorization and prediction using multi-omic data[J]. Comput Stat Data Anal, 2024, 197: 107974. doi:10.1016/j.csda.2024.107974

[38] Ghosal S, Chen Q, Pergola G, et al. A generative-discriminative framework that integrates imaging, genetic, and diagnosis into coupled low dimensional space[J]. Neuroimage, 2021, 238: 118200. doi:10.1016/j.neuroimage.2021.118200

[39] Han Y, Lam JCK, Li VOK, et al. Interpretable AI-driven causal inference to uncover the time-varying effects of PM2.5 and public health interventions on COVID-19 infection rates[J]. Humanit Soc Sci Commun, 2024, 11(1): 1713. doi:10.1057/s41599-024-04202-y

[40] Daunhawer I, Sutter TM, Marcinkevičs R, et al. Self-supervised disentanglement of modality-specific and shared factors improves multimodal generative models[J]. Pattern Recognition, 2021, 12544: 459-473. doi:10.1007/978-3-030-71278-5_33

[41] Argelaguet R, Arnol D, Bredikhin D, et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data[J]. Genome Biol, 2020, 21(1): 111. doi:10.1186/s13059-020-02015-1

[42] Shen R, Mo Q, Schultz N, et al. Integrative subtype discovery in glioblastoma using iCluster[J]. PLoS One, 2012, 7(4): e35236. doi:10.1371/journal.pone.0035236

[43] Xie H, Li J, Xue H. A survey of dimensionality reduction techniques based on random projection[EB/OL]. (2018-05-3)[2025-04-26]. https://doi.org/10.48550/arXiv.1706.04371

[44] Mirabnahrazam G, Ma D, Lee S, et al. Machine learning based multimodal neuroimaging genomics dementia score for predicting future conversion to Alzheimer's disease[J]. J Alzheimers Dis, 2022, 87(3): 1345-1365.

[45] Lock EF, Hoadley KA, Marron JS, et al. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types[J]. Ann Appl Stat, 2013, 7(1): 523-542.

[46] Yang Y, Ma C. Estimating shared subspace with AJIVE: the power and limitation of multiple data matrices[EB/OL]. (2025-02-15)[2025-04-26]. https://doi.org/10.48550/arXiv.2501.09336

[47] Gordon SL, Jahn E, Mazaheri B, et al. Identification of mixtures of discrete product distributions in near-optimal sample and time complexity[C]. Proceedings of the 37th Annual Conference on Learning Theory (COLT), PMLR, 2024: 2071-2091. https://proceedings.mlr.press/v247/gordon24a.html

[48] Yang J, Yu Y, Niu D, et al. ConFEDE: contrastive feature decomposition for multimodal sentiment analysis[C]. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023: 7617-7630.

[49] Freund MC, Etzel JA, Braver TS. Neural coding of cognitive control: the representational similarity analysis approach[J]. Trends Cogn Sci, 2021, 25(7): 622-638.

[50] Angelopoulos N, Chatzipli A, Nangalia J, et al. Bayesian networks elucidate complex genomic landscapes in cancer[J]. Commun Biol, 2022, 5(1): 306. doi:10.1038/s42003-022-03243-w

[51] Yan HX, Weng DW, Li DG, et al. Prior knowledge-guided multilevel graph neural network for tumor risk prediction and interpretation via multi-omics data integration[J]. Brief Bioinform, 2024, 25(3): bbae184. doi:10.1093/bib/bbae184

[52] Nelson Hayes C, Nakahara H, Ono A, et al. From omics to multi-omics: a review of advantages and tradeoffs[J]. Genes, 2024, 15(12): 1551. doi:10.3390/genes15121551

[53] Forés-Martos J, Forte A, García-Martínez J, et al. A trans-omics comparison reveals common gene expression strategies in four model organisms and exposes similarities and differences between them[J]. Cells, 2021, 10(2): 334. doi:10.3390/cells10020334

[54] Liu J, Cen X, Yi C, et al. Challenges in AI-driven biomedical multimodal data fusion and analysis[J]. Genomics Proteomics Bioinformatics, 2025, 23(1):

qzaf011. doi：10.1093/gpbjnl/qzaf011

[55] Jia Z, Giehl RFH, Meyer RC, et al. Natural variation of BSK3 tunes brassinosteroid signaling to regulate root foraging under low nitrogen[J]. Nat Commun, 2019, 10(1): 2378. doi：10.1038/s41467-019-10331-9

[56] Yip SS, Aerts HJ. Applications and limitations of radiomics[J]. Phys Med Biol, 2016, 61(13): R150-R166.

[57] Krishna A, Kurian NC, Patil A, et al. PathoGen-X：a cross-modal genomic feature trans-align network for enhanced survival prediction from histopathology images[C]. 2025 IEEE 22nd International Symposium on Biomedical Imaging, IEEE. doi：10.1109/ISBI60581.2025.10981028

[58] Yan Y, Yao XJ, Wang SH, et al. A survey of computer-aided tumor diagnosis based on convolutional neural network[J]. Biology (Basel), 2021, 10(11): 1084. doi：10.3390/biology10111084

[59] Butler L, Karabayir I, Samie Tootooni M, et al. Image and structured data analysis for prognostication of health outcomes in patients presenting to the ED during the COVID-19 pandemic[J]. Int J Med Inform, 2021, 158：104662. doi：10.1016/j.ijmedinf.2021.104662

[60] LI Y, Hajj HA, Conze PH, et al. Multimodal information fusion for the diagnosis of diabetic retinopathy[EB/OL]. (2023-03-20)[2025-04-26]. https://arxiv.org/abs/2304.00003

[61] Luo H, Huang JS, Ju HR, et al. Multimodal multi-instance evidence fusion neural networks for cancer survival prediction[J]. Sci Rep, 2025, 15(1): 10470. doi：10.1038/s41598-025-93770-3

[62] Li T, Zhou X, Xue J, et al. Cross-modal alignment and contrastive learning for enhanced cancer survival prediction[J]. Comput Methods Programs Biomed, 2025, 263：108633. doi：10.1016/j.cmpb.2025.108633

[63] Schneider L, Laiouar-Pedari S, Kuntz S, et al. Integration of deep learning-based image analysis and genomic data in cancer pathology：a systematic review[J]. Eur J Cancer, 2022, 160：80-91. doi：10.1016/j.ejca.2021.10.007

[64] Zheng T, Hu W, Wang H, et al. MRI-based texture analysis for preoperative prediction of BRAF V600E mutation in papillary thyroid carcinoma[J]. J Multidiscip Healthc, 2023, 16：1-10. doi：10.2147/JMDH.S393993

[65] Yu J, Ma T, Chen F, et al. Task-driven framework using large models for digital pathology[J]. Commun Biol, 2024, 7(1): 1619. doi：10.1038/s42003-024-07303-1

[66] Chen RJ, Lu MY, Williamson DFK, et al. Pan-cancer integrative histology-genomic analysis *via* multimodal deep learning[J]. Cancer Cell, 2022, 40(8): 865-878.

[67] Brussee S, Buzzanca G, Schrader AMR, et al. Graph neural networks in histopathology：emerging trends and future directions[J]. Med Image Anal, 2025, 101：103444. doi：10.1016/j.media.2024.103444

[68] Ding KX, Zhou M, Metaxas DN, et al. Pathology-and-genomics multimodal transformer for survival outcome prediction[M]//Medical Image Computing and Computer Assisted Intervention－MICCAI 2023. Cham：Springer Nature Switzerland, 2023：622-631. doi：10.1007/978-3-031-43987-2_60

[69] Qi YJ, Su GH, You C, et al. Radiomics in breast cancer：current advances and future directions[J]. Cell Rep Med, 2024, 5(9): 101719. doi：10.1016/j.xcrm.2024.101719

[70] Ehrenstein V, Kharrazi H, Lehmann H, et al. Obtaining data from electronic health records[EB/OL]. (2025-04-18)[2025-04-26]. https://www.ncbi.nlm.nih.gov/books/NBK551878/

[71] Patharkar A, Cai FL, Al-Hindawi F, et al. Predictive modeling of biomedical temporal data in healthcare applications：review and future directions[J]. Front Physiol, 2024, 15：1386760. doi：10.3389/fphys.2024.1386760

[72] Zhan X, Humbert-Droz M, Mukherjee P, et al. Structuring clinical text with AI：old versus new natural language processing techniques evaluated on eight common cardiovascular diseases[EB/OL]. (2025-04-18)[2025-04-26]. https://www.cell.com/patterns/fulltext/S2666-3899(21)00122-7

[73] Chen XL, Xie HR, Tao XH, et al. Artificial intelligence and multimodal data fusion for smart healthcare：topic modeling and bibliometrics[J]. Artif Intell Rev, 2024, 57(4): 91. doi：10.1007/s10462-024-10712-7

[74] Shajari S, Kuruvinashetti K, Komeili A, et al. The emergence of AI-based wearable sensors for digital health technology：a review[J]. Sensors, 2023, 23(23): 9498. doi：10.3390/s23239498

[75] Lih OS, Jahmunah V, Palmer EE, et al. EpilepsyNet：novel automated detection of epilepsy using transformer model with EEG signals from 121 patient population[J]. Comput Biol Med, 2023, 164：107312. doi：10.1016/j.compbiomed.2023.107312

[76] Deniz-Garcia A, Fabelo H, Rodriguez-Almeida AJ, et al. Quality, usability, and effectiveness of mHealth apps and the role of artificial intelligence：current scenario and challenges[J]. J Med Internet Res, 2023, 25：e44030. doi：10.2196/44030

[77] Basak H, Yin ZZ. Semi-supervised domain adaptive medical image segmentation through consistency

regularized disentangled contrastive learning[M]//Medical Image Computing and Computer Assisted Intervention-MICCAI 2023. Cham: Springer Nature Switzerland, 2023: 260-270. doi:10.1007/978-3-031-43901-8_25

[78] Zhao F, Zhang CC, Geng BC. Deep multimodal data fusion[J]. ACM Comput Surv, 2024, 56(9): 1-36.

[79] Li S, Tang H. Multimodal alignment and fusion: a survey[EB/OL]. (2024-11-26)[2025-04-26]. https://arxiv.org/abs/2411.17040

[80] Hangaragi S, Neelima N, Jegdic K, et al. Integrated fusion approach for multi-class heart disease classification through ECG and PCG signals with deep hybrid neural networks[J]. Sci Rep, 2025, 15(1): 8129. doi:10.1038/s41598-025-92395-w

[81] Domingo J, Minaeva M, Morris JA, et al. Non-linear transcriptional responses to gradual modulation of transcription factor dosage[J]. bioRxiv, 2024. doi: 10.1101/2024.03.01.582837

[82] Han GR, Goncharov A, Eryilmaz M, et al. Machine learning in point-of-care testing: innovations, challenges, and opportunities[J]. Nat Commun, 2025, 16(1): 3165. doi:10.1038/s41467-025-58527-6

[83] Kawahara D, Nagata Y. T1-weighted and T2-weighted MRI image synthesis with convolutional generative adversarial networks[J]. Rep Pract Oncol Radiother, 2021, 26(1): 35-42.

[84] Kang Z, He Y, Wang J, et al. Efficient multi-model fusion with adversarial complementary representation learning[EB/OL]. (2025-04-18)[2025-05-26]. https://ieeexplore.ieee.org/abstract/document/10650588/

[85] Yoon S, Byun S, Jung K. Multimodal speech emotion recognition using audio and text[EB/OL]. (2025-04-18)[2025-05-26]. https://ieeexplore.ieee.org/abstract/document/8639583/

[86] Höhn J, Krieghoff-Henning E, Jutzi TB, et al. Combining CNN-based histologic whole slide image analysis and patient data to improve skin cancer classification[J]. Eur J Cancer, 2021, 149: 94-101. doi:10.1016/j.ejca.2021.02.032

[87] Arnold C, Küpfer A. Alignment helps make the most of multimodal data[EB/OL]. (2024-05-14)[2025-04-26]. https://arxiv.org/abs/2405.08454

[88] Yang H, Zhou HY, Li C, et al. Multimodal self-supervised learning for lesion localization[EB/OL]. (2024-08-20)[2025-04-26]. https://ieeexplore.ieee.org/abstract/document/10635268/

[89] Lobato-Delgado B, Priego-Torres B, Sanchez-Morillo D. Combining molecular, imaging, and clinical data analysis for predicting cancer prognosis[J]. Cancers, 2022, 14(13): 3215. doi:10.3390/cancers14133215

[90] Chen RJ, Lu MY, Wang J, et al. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis[J]. IEEE Trans Med Imaging, 2022, 41(4): 757-770.

[91] Wijethilake N, Islam M, Ren HL. Radiogenomics model for overall survival prediction of glioblastoma[J]. Med Biol Eng Comput, 2020, 58(8): 1767-1777.

[92] Magbanua MJM, Li W, van't Veer LJ. Integrating imaging and circulating tumor DNA features for predicting patient outcomes[J]. Cancers (Basel), 2024, 16(10): 1879. doi:10.3390/cancers16101879

[93] Niu W, Yan J, Hao M, et al. MRI transformer deep learning and radiomics for predicting IDH wild type TERT promoter mutant gliomas[J]. NPJ Precis Oncol, 2025, 9(1): 89. doi:10.1038/s41698-025-00884-y

[94] Angelopoulos N, Chatzipli A, Nangalia J, et al. Bayesian networks elucidate complex genomic landscapes in cancer[J]. Commun Biol, 2022, 5(1): 306. doi:10.1038/s42003-022-03243-w

[95] Herawan M, Adriansjah R. Prostate specific antigen level and gleason score in Indonesian prostate cancer patients[EB/OL]. (2025-04-18)[2025-04-26]. https://repository.unar.ac.id/jspui/handle/123456789/8814

[96] Kabir Anaraki A, Ayati M, Kazemi F. Magnetic resonance imaging-based brain tumor grades classification and grading via convolutional neural networks and genetic algorithms[J]. Biocybern Biomed Eng, 2019, 39(1): 63-74.

(编辑:相峰)