面向乳腺癌 PAM50 亚型的多模型机器学习方法系统评估与生物信息学分析

Systematic Evaluation and Bioinformatics Analysis of Machine Learning Models for PAM50-Based Breast Cancer Subtyping

北京科技大学《数据科学：R 语言基础》大作业

组长：张鲴沣　北京交通大学
组员：刘陈子颖　中国地质大学
组员：文浩名　中国地质大学
组员：刘星雨　北京体育大学

日期：2025 年 6 月 21 日

摘要

本研究旨在对多种机器学习模型（LightGBM、XGBoost、多层感知器（MLP）、卷积神经网络（CNN）和支持向量机（SVM））在乳腺癌 PAM50 亚型分类任务中的应用进行系统性评估与比较分析。我们首先基于基因表达谱构建数据集，并采用五折交叉验证对各模型进行训练与测试。随后，结合网格搜索与随机搜索方法，对每种模型的关键超参数（如学习率、树的数量、隐藏层单元数及卷积核大小等）进行优化。实验结果表明，在准确率、精确率、召回率和 F1 值等指标上，LightGBM 与 XGBoost 表现出更佳的综合性能，而 CNN 在类别不平衡情况下展现出更强的泛化能力。最后，我们基于 SHAP（SHapley Additive exPlanations）方法对模型的特征重要性进行可视化分析，以揭示基因表达特征在不同 PAM50 亚型分类中的生物学意义。研究结论显示，集成学习模型在处理高维稀疏基因表达数据时效率更高，深度学习模型（CNN）虽然对数据量和计算资源需求较高，但在小样本不平衡场景下具有鲁棒性；基于 SHAP 的解释性分析进一步识别出若干关键基因在 Luminal A、Luminal B、HER2-enriched 和 Basal-like 亚型分类中的潜在作用。

关键词： 乳腺癌；PAM50 亚型；机器学习；LightGBM；XGBoost；CNN；SHAP；基因表达

Abstract

This study conducts a systematic evaluation and comparative analysis of various machine learning models—LightGBM, XGBoost, Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), and Support Vector Machine (SVM)—for PAM50-based breast cancer subtyping. Initially, we construct a dataset from gene expression profiles and apply five-fold cross-validation to train and evaluate each model. Next, a hybrid approach combining grid search and randomized search is employed to optimize key hyperparameters such as learning rate, number of trees, number of hidden units, and convolution filter sizes. Experimental results demonstrate that LightGBM and XGBoost achieve superior overall performance in terms of accuracy, precision, recall, and F1-score, whereas the CNN exhibits stronger robustness under class imbalance conditions. Finally, we utilize SHapley Additive exPlanations (SHAP) to visualize feature importance, revealing the biological relevance of gene expression features in distinguishing different PAM50 subtypes. The findings indicate that ensemble learning models handle high-dimensional sparse gene expression data more efficiently, deep learning models (CNN) require larger datasets and greater computational resources but perform robustly in small-sample, imbalanced scenarios, and SHAP-based interpretability analysis identifies several key genes implicated in Luminal A, Luminal B, HER2-enriched, and Basal-like subtype classification.

Keywords: breast cancer; PAM50 subtyping; machine learning; LightGBM; XGBoost; CNN; SHAP; gene expression

目录

1 引言

1.1 分子亚型视角下的乳腺癌异质性与分类演进

1.1.1 乳腺癌分类的早期演进：从病理形态学到 IHC 分型

乳腺癌作为一种临床常见的恶性肿瘤，其内在特征呈现出**显著的表型与遗传学差异**，构成了其**高度异质性的核心特质**。乳腺癌是全球女性中发病率最高的恶性肿瘤之一，其生物学行为呈现显著异质性，体现在组织学形态、基因组改变、转录组表达以及治疗反应等多个层面 [1]。

传统分型方法主要依赖免疫组织化学（IHC）评估雌激素受体（ER）、孕激素受体（PR）及 HER2 蛋白的表达。然而，IHC 方法存在主观性强、标准不一等问题，影响分型的准确性与一致性，且难以揭示亚型内部的分子异质性 [2]。

在精准医疗时代背景出现之前，乳腺癌的早期分类主要依赖于肿瘤组织的病理学形态观察以及关键受体（如雌激素受体 ER、孕激素受体 PR 及人表皮生长因子受体 2 HER2）的表达状态评估。基于免疫组织化学（IHC）检测的此类分型方法，将乳腺癌初步区分为激素受体阳性、HER2 阳性以及三阴性等主要临床亚组，这一分类策略在一定程度上揭示了不同乳腺癌亚组在生物学行为、疾病进展速度及治疗对策选择上的差异性，为临床决策制定提供了关键性依据。然而，仅凭借有限的临床病理特征组合，尚难以全面阐释乳腺癌内部更为精细与深邃的生物学多样性及其潜在的分子驱动机制。

1.1.2 高通量组学驱动的分子亚型发现与完善

分子生物学，特别是高通量组学技术的飞速发展，为实现更为精细化的肿瘤分子分型奠定了基础。具有里程碑意义的是，2000 年 Perou 及其合作者首次运用基因表达谱分析技术，前瞻性地提出了乳腺癌的 **“内在亚型”（intrinsic subtype）** 这一革命性概念，初步将乳腺癌划分为 Basal-like（基底样型）、HER2 富集型、Luminal A 型和 Luminal B 型等不同的分子亚型 [3]。

紧随其后，Sørlie 等通过更大规模的独立患者队列数据验证及长期临床随访研究，进一步确认并完善了这一分类体系，最终确立了五种核心的内在分子亚型，即 **Luminal A 型、Luminal B 型、HER2 过表达型（HER2-enriched）、Basal-like 型和 Normal-like 型** [4]。

这些基于全基因组转录谱定义的分子亚型，在核心驱动基因的表达模式、细胞信号通路激活状态、临床预后以及对特定治疗方案（如内分泌治疗、抗 HER2 靶向治疗）的敏感性等方面均展现出显著差异，深刻揭示了乳腺癌固有的分子异质性。例如，Luminal A/B 亚型通常表现为 ER 阳性且细胞分化程度相对较高，整体预后较为良好，尤其对内分泌治疗反应积极；与之相对，Basal-like 亚型（其特征与临床定义的三阴性乳腺癌高度重叠）则常表现出更强的侵袭性和转移潜能，预后普遍较差；而 HER2 过表达亚型则特异性地受益于针对 HER2 靶点的靶向治疗策略 [5]。

1.1.3 PAM50 模型的建立与多组学整合分型的深化

随着基因表达芯片技术的成熟以及新一代高通量测序（NGS）技术的广泛应用，科研人员得以对乳腺癌的分子景观进行更为深入和精细的剖析。在此背景下，Parker 等人基于精心筛选的 50 个关键基因（即 PAM50 基因集）的表达谱数据，成功构建了鲁棒的 **PAM50 分类模型**。该模型能够将待检肿瘤样本精确地归类至上述五种内在亚型之一，并已被证实可有效辅助临床治疗决策的制定，如预测化疗敏感性和复发风险等 [5]。PAM50 分类器的出现及其后续在临床实践中的推广应用，标志着乳腺癌分子分型策略从基础研究向临床实践转化的重要里程碑。

与此同时，国际大型癌症基因组协作研究项目（例如 The Cancer Genome Atlas, TCGA；以及 Molecular Taxonomy of Breast Cancer International Consortium, METABRIC 等）通过整合数千例乳腺癌患者的基因组、转录组等多维度组学数据，进行了迄今为止最为全面的综合性分子分析，其成果不仅验证了经典分子亚型的存在，更进一步鉴定出比传统四分类或五分类更为精细的分子亚群。例如，Curtis 及其团队在对超过 2000 例乳腺肿瘤样本的基因组和转录组数据进行深度挖掘后，创新性地提出了包含 **10 个整合性分子分型簇（Integrative Clusters, IntClusts）**的新分类框架，其中每一个簇均展现出独特的基因组变异特征组合、关键信号通路激活模式以及与之相关的临床结局差异，从而揭示了乳腺癌异质性更为复杂和精细的分子图谱 [6]。

这些前沿研究共同表明，即使在传统的同一分子亚型内部，仍可能蕴含进一步的转录组学差异及相应的生物学行为变异，提示乳腺癌的分子异质性远超早期认知。在精准医学背景下，分子分型已成为乳腺癌个体化治疗的关键基础。不同分子亚型对靶向治疗的反应存在显著差异：携带 BRCA1/2 突变的患者对 PARP 抑制剂（如 olaparib）敏感 [7]，而 HER2-enriched 亚型可显著获益于抗 HER2 药物治疗（如 trastuzumab）[8]。近年来提出的 HER2-low 亚型（IHC 1+ 或 IHC 2+/ISH-）被认为具有独特分子特征，如 PIK3CA 突变频率升高、EGFR 扩增趋势等，提示其可能对 PI3K/Akt 通路抑制剂敏感 [9]。这些进展不仅深化了对乳腺癌发病机制的理解，也为临床精准治疗提供了新路径。

1.1.4 三阴性乳腺癌（TNBC）的精细分子分型与 TCGA 数据资源

三阴性乳腺癌（TNBC）是乳腺癌高度异质性的一个典型范例。传统临床定义上，TNBC 指 ER、PR 及 HER2 表达均为阴性的乳腺癌亚组，但分子层面的深入研究清晰表明，TNBC 并非一个均质化的单一疾病实体。

Lehmann 等人率先基于基因表达谱特征，将 TNBC 细致地区分为 Basal-like 1 (BL1)、Basal-like 2 (BL2)、间充质样 (Mesenchymal, M)、免疫调节型 (Immunomodulatory, IM) 以及雄激素受体阳性 (Luminal Androgen Receptor, LAR) 等多个具有不同生物学特性和潜在治疗靶点的分子亚型。随后，Burstein 等学者通过整合 RNA 表达谱和 DNA 水平的基因组变异分析，进一步确认并优化了 TNBC 的分子分型，提出了包含四个更为稳定且临床意义明确的分子亚型：**雄激素受体阳性型（LAR）、间充质型（MES）、基底样免疫抑制型（Basal-**

like Immune-Suppressed, BLIS）和基底样免疫激活型（Basal-like Immune-Activated, BLIA）。研究证实，这四种 TNBC 亚型在驱动基因、信号通路、肿瘤微环境特征以及对标准化疗和新兴靶向治疗（如 PARP 抑制剂、免疫检查点抑制剂）的临床反应方面均存在显著差异 [10]。

此分类体系的演进进一步凸显了乳腺癌内在生物学行为的复杂性与多样性，并为 TNBC 的个体化治疗策略开发提供了重要理论依据。在乳腺癌分子分型的标准化与机制研究方面，美国癌症基因组图谱（The Cancer Genome Atlas, TCGA）乳腺癌项目（TCGA-BRCA）提供了关键资源。该项目整合了超过 1,000 例乳腺癌样本的多组学数据，包括基因突变、拷贝数变异、RNA 表达、miRNA 表达、DNA 甲基化及蛋白质组学信息，并配有详尽的临床注释（如 PAM50 分型、预后、治疗方案等）[1]。Ciriello 等人基于 TCGA-BRCA 数据构建了乳腺癌的综合基因组分类图谱，揭示了 HER2-enriched 亚型中广泛的 ERBB2 基因扩增现象 [11]。该数据集凭借其高质量的多模态特征和标准化处理流程，已成为分型模型开发、生物标志物筛选与机制研究的核心平台。

1.1.5 分子分型对临床实践的意义与未来挑战

乳腺癌分类体系从最初依赖组织学形态观察，到逐步引入基于关键蛋白表达的免疫组化分型，再到当前广泛应用的基于全基因组表达谱的分子分型，这一演进历程极大地深化了科学界对乳腺癌内在异质性的认知。分子亚型分类不仅精确揭示了不同肿瘤亚群在核心发病机制、疾病自然史及临床预后方面的固有差异，更为实现基于分子特征的个体化治疗策略奠定了坚实基础。例如，针对 HER2 基因扩增或过表达的 HER2 阳性亚型乳腺癌患者，曲妥珠单抗、帕妥珠单抗等一系列高效靶向药物的成功研发与应用，显著改善了其临床结局；而对于 ER 阳性的 Luminal 型乳腺癌，以内分泌治疗为主的综合治疗方案则取得了显著疗效。这些辉煌的研究进展共同构筑了利用分子特征对乳腺癌进行精细化分类的理论与实践框架。

然而，随着多组学数据的爆炸式增长和对肿瘤复杂性认识的不断深化，新的挑战亦随之而来：**如何有效地整合日益复杂且海量的高维多组学数据，以进一步提升分子分型的准确性、稳定性及其生物学意义与临床转化价值**，是未来研究亟待解决的关键科学问题。

1.2 基因组数据背景：在乳腺癌 PAM50 亚型分类中的应用

1.2.1 PAM50 基因表达谱：临床与生物学意义

PAM50 基因表达谱是一种包含 50 个基因的分子检测方法，广泛用于将乳腺癌分为五个内在分子亚型：Luminal A、Luminal B、HER2 富集型、基底细胞样型和正常乳腺样型。这些亚型不仅在预后上存在显著差异——例如，Luminal A 型通常预后最好，而 HER2 富集型和基底细胞样型则更具侵袭性——而且在潜在的生物学驱动机制上也各不相同 [21]。PAM50 检测可以通过微阵列、RNA 测序（RNASeq）或定量逆转录聚合酶链反应（qRT-PCR）等技术平台获取的基因表达数据来进行 [21]，并已作为一种预后工具应用于临床实践 [22]。

然而，PAM50 分类并非没有挑战。研究表明，基于 PAM50 的分子分型与基于免疫组化（IHC）的替代分型方法之间可能存在不一致性 [21]。此外，基因表达数据的预处理方法，例如改良的中位数基因中心化（MMGC）或亚组特异性基因中心化（SSGC），都可能影响 PAM50 亚型的分配结果 [21]。这提示我们，在解读基于 PAM50 的机器学习模型结果时，需要考虑这些潜在的变异来源。更有趣的是，PAM50 亚型在肿瘤进展过程中可能发生转变，例如从原发灶到转移灶，通常趋向于更具侵袭性的亚型，这对患者的预后具有重要影响 [23]。虽然本研究主要关注原发肿瘤的分类，但亚型的动态演变是未来研究中不容忽视的背景因素。

在 PAM50 的五个亚型中，“正常乳腺样型”（Normal-like）的存在为机器学习模型的构建和解读带来了一定的复杂性。该亚型可能代表了混杂在肿瘤样本中的正常乳腺组织，也可能是一种具有独特生物学特征的肿瘤实体 [21]。如果模型训练数据中包含了“正常乳腺样型”这一类别，那么如何区分其与真实的正常组织样本（如果后者也被纳入训练或作为对照）以及其他肿瘤亚型，将直接影响模型的性能指标和生物学解释的准确性。本研究的实验设置中关于如何处理“正常乳腺样型”样本，或者是否将正常样本作为一个独立的类别纳入训练的细节，对于全面理解模型性能至关重要。因此，在评估模型时，应考虑到“正常乳腺样型”的模糊性及其处理方式对模型性能和生物学结论的潜在影响。

1.2.2 TCGA-BRCA：乳腺癌基因组学的重要资源

癌症基因组图谱（The Cancer Genome Atlas, TCGA）项目旨在通过大规模基因组测序和多维度综合分析，系统地编目和发现主要的致癌基因组变异，为超过 30 种人类肿瘤（包括乳腺癌，即 TCGA-BRCA）创建全面的基因组图谱 [24]。TCGA 的数据是公开的，为科研人员理解肿瘤发生、改进诊断方法和治疗标准提供了宝贵的资源 [24]。

TCGA-BRCA 数据集已被广泛应用于各种基于机器学习的研究中。例如，有研究利用 TCGA-BRCA 数据（分析了 750 名患者及其 PAM50 亚型信息）构建了基于 TP53 突变状态的基因共表达网络 [25]。其他研究则利用其基因表达数据进行癌症类型或亚型的分类 [26]。这些应用充分证明了 TCGA-BRCA 数据的丰富性和在机器学习驱动的癌症研究中的广泛适用性，它很可能是本研究所用数据集的来源（或类似来源）。

尽管 TCGA 数据非常全面，但由于其数据来源于多个中心，且可能是在不同时间段收集的，因此潜在的批次效应（Batch Effects）是一个需要关注的问题。如果未能妥善处理，批次效应可能会掩盖真实的生物学信号，从而干扰分析结果。[26] 的研究在整合 TCGA 和 GTEx 数据时，明确使用了“recount pipeline”来移除批次效应。本研究在处理 TCGA 衍生数据时，对批次效应的校正方法（若有）对模型性能有显著影响。如果数据确实来源于 TCGA 且未详细说明批次校正过程，那么这应被视为一个潜在的变数。不同研究对批次效应的不同处理方式，也可能导致模型性能的差异，或限制了直接比较的可靠性。

1.2.3 癌症中的基因拷贝数（GCN）分析

基因拷贝数变异（Copy Number Variation, CNV），或称基因拷贝数（Gene Copy Number, GCN），是癌症基因组中的一类重要变异。GISTIC2 (Genomic Identification of Significant Targets in Cancer 2.0) 是一种广泛应用的算法，用于在大量样本中识别显著扩增或缺失的基因组区域。该算法为每个畸变区域分配一个 G-score（综合考虑畸变的幅度和在样本间的发生频率），并计算错误发现率（False Discovery Rate, FDR）q 值，以确定统计学上显著的区域 [27]。GISTIC2 的输出包括畸变区域的峰值区（Peak Region）和宽峰区（Wide Peak），后者对确定区域内最可能的靶基因更为稳健。其输出文件中的 `Seg.CN` 列通常表示拷贝数的 log2 比率，而数值“1”或“2”可以分别指示低水平或高水平的拷贝数畸变 [27]。

一个来自 TCGA 胃腺癌（STAD）的 GISTIC2 分析实例展示了通过该方法可以识别出许多已知的癌基因和抑癌基因区域的显著扩增（如 ERBB2, CCNE1, KRAS, MYC, CCND1, EGFR, FGFR2）或缺失。这突显了 GCN 数据在揭示癌症驱动因素方面的生物学相关性。

需要注意的是，“GCN”这一缩写在文献中可能有多种含义。除了指基因拷贝数外，它也可以指基因共表达网络（Gene Correlation Networks），如 [25] 中基于 TP53 突变状态为 TCGA-BRCA 患者构建的网络，该网络使用的是基因表达和生存数据。此外，“GCN”还可以指图卷积神经网络（Graph Convolutional Networks/Neural Networks, GCNNs），这是一种用于处理图结构数据的深度学习模型，可整合基因表达数据与蛋白质相互作用（PPI）网络或共表达网络等信息，以进行癌症亚型分类 [28]。在本报告的上下文中，当提及“GCN 数据”时，主要指的是基因拷贝数数据及其分析。

尽管本研究目前主要侧重于利用基因表达数据进行 PAM50 亚型分类，但基因拷贝数变异是癌症发生发展的基础分子事件之一，并常常驱动基因表达的改变。[27] 和相关 GISTIC2 分析结果（例如，ERBB2, MYC, CCNE1 等基因的扩增）直接关联到特定癌症亚型的驱动机制（例如，ERBB2 扩增是 HER2 富集型乳腺癌的标志）。[37] 中描述的 CopyClust 算法明确使用拷贝数数据进行乳腺癌的整合亚型分类。这表明，将拷贝数数据与基因表达数据整合，有可能提供一个更稳健、生物学意义更明确的分类框架，正如 CopyClust 等工具和多组学整合的大趋势（ [29]）所倡导的那样。一个值得深入探讨的问题是，当前基于表达谱的模型所产生的错误分类是否与特定的拷贝数谱相关联？这提示我们，未来的研究应考虑将 GCN（拷贝数）数据与现有模型所使用的基因表达数据进行整合的潜在益处和具体方法。

1.3 面向乳腺癌亚型分类的机器学习方法：算法进展与实践挑战

1.3.1 机器学习在乳腺癌亚型分类中的早期应用与进展

随着生物信息学技术的日新月异以及机器学习（Machine Learning, ML）理论与算法的持续发展，研究人员日益倾向于运用先进的计算方法，以期实现自动化、高精度地判别乳腺癌的分子亚型。纵观其发展历程，从早期的基于探索性数据分析的聚类算法和相对简单的传统分类器，到近年来结构更为复杂、功能更为强大的现代机器学习模型（尤其是深度学习模

型）的引入，相关算法在乳腺癌亚型分类任务中的应用取得了显著进展与广泛应用。

最初的乳腺癌内在亚型分类方法（例如经典的 PAM50 分类器），其核心机制可以被理解为一种基于样本间基因表达谱距离（如欧氏距离或相关性距离）计算的最近邻（Nearest Neighbor）或原型（Prototype-based, 如 Nearest Centroid）分类策略的简化机器学习模型 [5]。在此基础上，一系列经典的监督学习算法，如**支持向量机（SVM）、随机森林（Random Forest, RF）、k-近邻（kNN）**等，被广泛应用于处理高维基因表达数据，以实现对不同分子亚型的有效区分 [12]。

例如，Wu 等人利用 TCGA 数据库中的大规模乳腺癌转录组数据，系统地训练并比较了包括 SVM、kNN、朴素贝叶斯（Naive Bayes）和决策树（Decision Tree）在内的多种传统机器学习模型，用于区分三阴性与非三阴性乳腺癌患者。研究结果显示，在该特定分类任务中，SVM 模型展现出最高的分类准确率 [12]。此类研究初步证实了传统机器学习方法在从高维基因表达谱中提取有效判别信息，并将其应用于乳腺癌亚型判别任务中的可行性与潜力。近年来，研究者广泛利用 TCGA-BRCA 数据结合机器学习和深度学习方法进行乳腺癌分型建模。例如，Tong 等人提出基于图神经网络（GNN）的多组学融合模型，在亚型识别上优于传统方法 [13]。这些算法表明深度模型在揭示复杂生物学模式方面具有巨大潜力。

1.3.2 创新机器学习算法的涌现与深度学习的引入

近年来，针对乳腺癌特定亚型分类问题而专门设计的创新算法不断涌现，这些算法不仅在模型结构上有所突破，更在特征表示、信息融合等方面进行了深度探索，充分展现出方法学层面持续的多样化与复杂化趋势 [14–17]。例如，Rhee 等人创新性地提出将图卷积网络（Graph Convolutional Network, GCN）与关系网络（Relation Network）相结合的混合模型架构，通过有效整合基因表达谱数据与已知的蛋白质相互作用（Protein-Protein Interaction, PPI）网络信息，以期从系统生物学的角度改进乳腺癌亚型的分类精度与生物学解释性 [14]。

Gao 等学者则设计并实现了一种名为 DeepCC 的深度学习框架，该算法首先基于基因表达谱数据计算每个样本在预定义生物学通路上的富集分数，随后将这些通路活性谱作为输入特征，馈入一个全连接神经网络（Fully Connected Neural Network）进行训练，从而实现对癌症分子亚型的精确分类 [15]。

Beykikhoshk 及其团队则巧妙地运用了注意力机制（Attention Mechanism），针对 Luminal A 型和 Luminal B 型乳腺癌这两个临床上易于混淆的亚型，为每个待测样本动态计算出一组个性化的生物标志物及其相应的权重得分，进而依据这些加权特征进行亚型判别，有效提升了分类的精准度与可解释性 [16]。Lee 等人则进一步拓展了图神经网络的应用，他们开发了一种基于生物学通路信息的多重注意力图卷积网络模型（Pathway-associated Graph Attention Network），该模型在多个独立的乳腺癌数据集上均展现出稳定且优异的亚型分类性能，证实了其良好的泛化能力 [17]。

此外，Yu 等研究者针对乳腺癌不同分子亚型间的复杂性和重叠性问题，提出了一种“一对余”（One-vs-Rest）的策略，即针对每一种特定的分子亚型，首先筛选出与其显著相关的差

异表达基因集，然后基于这些特异性基因集分别训练二分类模型，旨在提高对那些特征边界较为模糊或易于混淆的亚型的辨识能力与分类鲁棒性 [12]。上述方法充分发掘了机器学习在处理高维基因组数据及复杂模式识别方面的潜力，持续推动分类准确率的提升。

1.3.3 机器学习在乳腺癌亚型分类中面临的挑战与展望

尽管机器学习方法在乳腺癌亚型分类研究中取得了令人鼓舞的进展，然而在实际应用与临床转化过程中仍面临诸多挑战。这些挑战不仅关乎模型本身的性能，更涉及到生物医学数据的固有特性与临床实践需求之间的契合度。具体而言，主要问题集中在以下几个方面：

- 首先是乳腺癌分子数据（尤其是基因组及转录组数据）普遍存在的**高维度、低样本量（High-Dimension, Low Sample Size, HDLSS）**特性，即特征数量远超样本数量（$p \gg n$）。此种数据结构极易导致所构建模型易于陷入过拟合状态，其泛化能力在独立验证数据集上往往不尽如人意，难以稳定推广至新的临床样本 [18,19]。
- 其次，来源于不同研究队列、采用不同实验技术平台或数据预处理流程所产生的组学数据，往往存在显著的**批次效应（batch effects）和固有技术平台差异**。直接整合此类异质性数据集以训练统一模型，常因这些非生物学因素的干扰而影响模型的稳定性、可重复性及最终的预测精度。
- 对于监督学习模型而言，**亚型标签（ground truth）的获取本身亦存在一定程度的不确定性与潜在偏差**：大多数临床样本的分子亚型标签是通过基因表达谱检测（如 PAM50 检测）或免疫组化（IHC）染色等方法间接推断得出，而不同的检测平台、抗体选择、判读标准以及阈值设定等因素均可能导致亚型标签在不同研究间甚至同一研究内部存在不一致性。这种标签噪声无疑为监督学习模型的稳健训练带来了严峻挑战。
- 虽然支持向量机、随机森林等传统机器学习模型在处理中低维度或样本量相对充足的数据时表现良好，但当特征维度急剧增加至远超样本规模时，其决策边界的构建可能受到高维空间中噪声特征的显著干扰，导致性能下降。因此，如何实施**有效的特征选择、特征提取、数据降维或模型正则化策略**，以从海量原始特征中精准提炼出真正与亚型相关的、具有生物学意义的核心信号，并抑制噪声影响，构成了模型构建流程中至关重要的环节之一。
- **模型的可解释性（interpretability）也是一个亟待解决的重大挑战**：临床医生和生物学研究者不仅关注模型预测的准确性，更渴望理解模型做出特定亚型判别的生物学依据，例如哪些特定的基因、通路或分子特征在驱动分类决策中起到了关键作用。然而，许多高性能的机器学习模型，特别是结构复杂的深度神经网络，其内部决策机制往往如同一个“黑箱”，难以直接阐释，亟需借助额外的可解释性分析方法（如 SHAP、LIME 等）以提升模型的透明度与临床可信度 [20]。

以上这些错综复杂的挑战，在一定程度上制约了机器学习模型在乳腺癌亚型分类领域向临床实践的有效转化与广泛应用。但机器学习在乳腺癌亚型精准分类中的应用前景依然广阔

且充满潜力。一方面，其强大的数据整合与模式识别能力，有助于突破传统基于单一或少数生物标志物进行分类的局限性，从而更全面地刻画乳腺癌的分子异质性。另一方面，机器学习亦为探索和发现潜在的新型分子亚型或数据中隐藏的复杂生物学模式提供了强大的计算工具。

随着相关算法的持续改进、多组学数据的不断积累以及对肿瘤生物学认识的日益深化，**机器学习有望在推动乳腺癌分子分型的精确化、个体化诊疗策略的制定以及新型治疗靶点的发现等方面发挥愈加 pivotal 的作用**。

1.4 多组学整合与深度学习：复杂生物学特征的建模新策略

1.4.1 多组学整合分析的必要性与挑战

乳腺癌的发生与演进是一个多阶段、多因素参与的复杂生物学过程，其分子基础涉及基因组（如基因突变、拷贝数变异、染色体结构异常）、转录组（如 mRNA、非编码 RNA 的表达谱改变）、表观基因组（如 DNA 甲基化、组蛋白修饰）、蛋白质组（如蛋白质表达、翻译后修饰及相互作用网络）乃至代谢组等多个分子层面的协同失调与异常累积 [30]。

传统的基于单一组学（single-omics）数据的分析策略，虽然能够在特定分子层面揭示部分生物学信息，但往往难以全面捕捉不同分子层级间复杂的相互作用网络与协同调控机制，从而限制了对肿瘤整体生物学特性的认知深度。鉴于此，**多组学整合分析（multi-omics data integration）**应运而生，旨在从系统生物学的视角整合来源于同一组患者样本的异质性的分子数据，以期揭示肿瘤发生发展的复杂生物学全貌，并从中发掘更为精准和鲁棒的生物标志物。

例如，对于同一批肿瘤样本，研究人员可以同步检测其 DNA 序列变异及拷贝数改变、全基因组 DNA 甲基化谱、mRNA 及 miRNA 等非编码 RNA 的表达丰度、以及关键蛋白质的表达水平与磷酸化状态等信息。这些多维度数据共同为每个肿瘤样本构建了更为全面且个体化的“分子指纹图谱”，然而其内在的高维度特性、不同组学数据间的异质性（如数据类型、尺度、分布差异）以及潜在的数据缺失问题，也对后续的计算分析方法提出了更高的要求与挑战。

1.4.2 深度学习在多组学整合中的应用与代表性模型

近年来，以**深度学习（Deep Learning）**为代表的人工智能技术取得了突破性进展，并迅速渗透至生物医学研究的各个领域，为复杂高维多组学数据的有效整合与深度特征学习开辟了新的策略途径 [31, 32]。

相较于传统的机器学习算法往往需要依赖人工设计的特征工程步骤，深度学习模型（例如深度神经网络 DNN、卷积神经网络 CNN、循环神经网络 RNN、图神经网络 GNN 以及自编码器 AE 等）具备从原始数据中自主学习多层次抽象特征表示的卓越能力，从而能够捕捉潜藏于数据中的线性和非线性复杂模式。这一特性使其高度契合于处理多组学数据中不同

分子层级间固有的复杂关联与调控关系。例如，Chaudhary 等学者成功开发了一种基于深度神经网络的多组学整合模型，该模型有效融合了肝癌患者的基因表达谱、DNA 甲基化谱以及 miRNA 表达谱数据，实验结果表明，相较于基于单一组学数据的模型，该整合模型能够更准确地预测患者的生存结局 [33]。

在乳腺癌亚型分类这一具体研究方向上，针对乳腺癌亚型分类任务的多组学深度学习模型亦在近年来逐渐涌现并展现出良好潜力。

近期，Choi 等人设计并提出了一种名为“**moBRCA-net**”的多组学整合深度学习框架，专门用于乳腺癌分子亚型的精准分类 [33]。该框架巧妙地整合了来自同一患者队列的基因表达数据、DNA 甲基化数据以及 miRNA 表达数据这三种关键的组学信息。其核心创新在于为每一种组学数据分别配备了一个自注意力机制（self-attention mechanism）模块，该模块能够动态地学习不同分子特征（如基因、甲基化位点、miRNA）在亚型分类决策过程中的相对重要性或贡献权重，从而在多模态信息融合过程中能够自适应地突出最具判别力的关键生物学信号，并抑制无关噪声的干扰 [33]。

通过在公开数据集上的严格评估，moBRCA-net 框架显示出，相较于仅依赖单一组学数据训练的基线模型，多组学数据的有效融合显著提高了乳腺癌亚型预测的准确性与鲁棒性。与之类似，Wang 及其合作者开发了一种名为 **MOGONET (Multi-Omics Graph cOnvolutional NETworks)** 的通用型多组学整合模型 [34]。MOGONET 利用图卷积网络（GCN）对每一种组学数据分别进行特征表示学习，捕获其内部的复杂结构信息；同时，通过设计跨组学的关联学习机制（如 view correlation discovery network），将从不同组学数据中学习到的特征表示进行有效汇总与对齐。该模型不仅成功应用于包括乳腺癌在内的多种肿瘤类型的多类别分子分型任务，且在多项生物医学分类基准测试中均表现出优于既往方法的性能。这些前沿研究实例充分证明，相较于传统上主要依赖基因表达谱等单一数据源的分类策略，精心设计并有效融合多组学数据的深度学习模型能够更全面、更深入地表征肿瘤的内在生物学特性，进而显著提升了分类模型的整体性能。

1.4.3 多组学深度学习方法优势与挑战

多组学深度学习方法应用于乳腺癌亚型分类任务，展现出诸多显著优势：

- 首先，通过同步考量来源于不同分子层面的信息（例如，特定基因突变如何驱动下游基因表达模式的改变，或者表观遗传修饰如何精细调控关键基因的转录活性等），整合模型能够捕捉到各亚型之间更为全面且细微的分子差异，从而有效提高最终分类结果的生物学相关性与临床意义 [30]。
- 深度学习模型所具备的强大非线性拟合能力，使其有潜力发掘出传统线性统计方法或浅层学习模型难以有效察觉的复杂高阶交互模式。例如，某些亚型特异性的分子特征组合或协同调控网络，可能只有在联合分析基因组变异、表观遗传修饰与转录组表达数据时才能清晰显现，而深度学习模型则有望自动从数据中挖掘并学习这类复杂的交互效应。

- 再次，深度学习模型（尤其是基于自编码器等架构的模型）可通过构建层次化的特征表示学习机制，将原始的高维、稀疏多组学数据逐步抽象并压缩至维度更低但信息含量更密集的判别性特征空间，从而在一定程度上有效缓解由数据高维性引发的“维度诅咒”问题，进而有效降低数据固有噪声及冗余信息对于模型性能的干扰。
- 随着自编码器（Autoencoders）、注意力机制（Attention Mechanisms）以及图神经网络（Graph Neural Networks）等先进技术的引入与融合，现代多组学深度学习模型在自动提炼关键判别特征以及在一定程度上提升模型决策过程可解释性方面均取得了积极进展。例如，注意力机制不仅能够提升模型的预测性能，其所学习到的注意力权重分布还可以直观地指示不同组学特征或样本区域在判别特定亚型时的相对重要性，为研究者识别潜在的生物标志物或理解模型决策逻辑提供了有价值的线索 [17]。

当然，将多组学深度学习方法应用于乳腺癌亚型分类等复杂生物医学问题，并非一帆风顺，仍面临一系列不容忽视的挑战。**不同组学数据间的规模不一致性与异质性问题**（例如，基因表达数据通常较为完整，而蛋白质组或代谢组数据可能存在较多缺失值或样本覆盖不全）；**如何有效处理数据缺失值并进行跨模态数据对齐**；**模型训练所需计算资源较高以及对大规模高质量标注数据的依赖性**；以及尽管有所改进但仍相对复杂的**模型可解释性问题**等。

目前现有多模态集成方法，如多核学习、典型相关性分析（CCA）等，往往依赖人工特征选择或仅能捕获线性关联，缺乏自动提取和融合不同组学数据中复杂非线性关系的能力 [35]，这限制了其在深度挖掘生物学机制方面的潜力。总体而言，**整合多源异质性数据的深度学习策略无疑为乳腺癌亚型分类研究提供了更为全面深入的分析视角和更为强大的计算建模工具**。在已公开发表的相关研究中，精心设计的多组学整合模型往往在分类准确度、模型鲁棒性以及对新样本的泛化能力等方面，均较传统的单组学模型展现出显著的优越性 [17]。

通过进一步优化深度学习模型的网络架构设计、创新多模态数据融合机制、以及更紧密地结合领域生物学知识对模型进行有效约束与引导，我们有望从海量复杂的多组学数据中精准鉴别出驱动不同乳腺癌亚型发生发展的关键分子特征与核心调控网络，从而有望将乳腺癌的精细化分子分型研究推向新的高度，并为实现真正的个体化精准医疗提供有力支撑。

1.5 高维低样本问题（HDLSS）与数据增强技术：泛化能力与生物学信号提取

1.5.1 高维低样本量（HDLSS）问题：维度诅咒与挑战

在生物医学数据分析领域，特别是涉及基因组、转录组、表观基因组等多组学数据的研究中，**高维度、低样本量（High-Dimension, Low Sample Size, HDLSS）**现象构成了一项普遍存在且极具挑战性的难题。在乳腺癌多组学研究的背景下，此问题表现得尤为突出与棘手。

典型的转录组测序数据或全基因组 DNA 甲基化芯片数据，其特征维度（例如，检测到的基因数量或 CpG 位点数量）往往高达数万甚至数十万级别，然而受限于临床样本获取的难度、研究成本以及伦理考量等因素，可供分析的患者样本数量（即观测值数量）通常仅有

数十到数百例。

这种“变量数量远超观测值数量”（$p \gg n$）的极端不平衡数据结构，极易导致机器学习模型在训练过程中过度拟合训练数据中的特有噪声和随机波动，而非学习到具有普适性的真实生物学规律。其直接后果便是，模型在训练集上表现优异，但在未见数据上泛化能力欠佳，即遭遇了所谓的**“维度诅咒”（curse of dimensionality）** [18]。因此，在 HDLSS 情境下，如何有效提升模型的泛化性能并从中提炼出稳健可靠的生物学信号，构成了此类研究的核心议题。

1.5.2 特征选择与数据降维策略

面对高维数据的挑战，一种直接且广泛采用的策略是在构建预测模型之前，对原始特征集进行审慎的筛选与精简，旨在通过剔除冗余及不相关特征，有效降低数据维度，从而聚焦于更具信息量的生物学信号，同时减轻模型的计算负担并可能提升其可解释性。

在过去的几十年中，研究人员针对高维基因表达数据等生物医学数据开发了大量的特征选择（feature selection）与特征提取（feature extraction）方法。这些方法大致可分为过滤式（filter）、包裹式（wrapper）和嵌入式（embedded）三大类。常见的实例包括：基于统计学检验（如 t 检验、ANOVA、卡方检验等）的单变量过滤方法，用以初步筛选出在不同亚型间表达水平存在显著差异的基因；基于稀疏学习理论的嵌入式方法，如 LASSO（Least Absolute Shrinkage and Selection Operator）回归或岭回归（Ridge Regression）及其变体（如 Elastic Net），它们通过在损失函数中引入 L1 或 L2 范数惩罚项，能够在模型训练过程中自动选择重要特征并将其余特征的系数压缩至零或接近零；以及无监督的数据降维方法，如主成分分析（Principal Component Analysis, PCA）、t-分布随机邻域嵌入（t-SNE）或均匀流形逼近与投影（UMAP），**它们旨在将原始高维数据投影到较低维度的子空间，同时尽可能保留数据的原有结构或方差信息** [18]。

实践证明，在乳腺癌亚型分类任务中，恰当的特征选择或降维预处理步骤，往往能够显著提升后续分类模型的性能、稳定性以及最终结果的生物学可解释性。例如，广为人知的 PAM50 基因集本身即是从数千个候选基因中通过严格筛选得到的 50 个代表性基因，利用这些预选的标志性基因而非全基因组表达谱来训练分类模型，不仅能有效减少数据噪声的干扰，增强模型的生物学意义，还能降低后续临床检测的成本与复杂性。

1.5.3 数据增强技术：生成模型（GANs, VAEs）的应用

除了从特征层面入手应对 HDLSS 问题外，另一个旨在提升模型泛化能力的重要策略是**数据增强（data augmentation）**，即通过各种技术手段在维持数据原有分布特性的前提下，“人为地”扩充训练数据集的有效样本规模，从而为模型提供更丰富、更多样化的学习素材。

在计算机视觉领域，基于图像旋转、翻转、裁剪、色彩抖动等几何或光学变换的数据增强技术已被广泛证明能够显著提高深度学习模型的鲁棒性与泛化能力。然而，对于基因组表

达谱、DNA 甲基化谱等结构化的数值型高维数据，传统的基于简单几何变换或噪声注入的数据增强方法对于此类结构化数值型数据往往不直接适用或效果有限。

近年来，以**生成对抗网络（Generative Adversarial Networks, GANs）**为代表的深度生成模型（deep generative models）的兴起，为此类数据的增强任务提供了富有潜力的新思路。GAN 由一个生成器（Generator）和一个判别器（Discriminator）组成，两者通过一种对抗性的“零和博弈”过程进行迭代训练：生成器致力于学习真实数据的内在分布并产生尽可能逼真的“伪样本”，而判别器则努力区分真实样本与生成器产生的伪样本。通过这种动态的竞争与协同进化，理想状态下生成器能够捕捉到真实数据的高维复杂分布，并据此生成具有高度真实感和多样性的新合成样本。研究人员已积极尝试将 GAN 及其变体应用于基因表达数据的增强任务中：例如，Chaudhari 等人提出了一种改进型的 GAN 架构（**MG-GAN**），其在生成器的网络结构中引入了高斯噪声层进行扰动，实验证明该方法能够成功模拟出与真实癌症基因表达数据分布高度相似的合成数据 [31]。

Kwon 等学者在实践中发现，直接利用包含数万基因的完整表达矩阵来训练 GAN 模型，其效果往往不尽如人意，合成数据的质量难以保证。为此，他们提出了一种策略，即首先通过特征选择方法筛选出一个信息量较为集中的显著基因子集，然后仅针对这个选定的基因子集（而非全基因组）进行 GAN 模型的训练与数据生成，从而有效提高了合成数据的质量与生物学相关性 [32]。

Ahmed 等研究者则进一步将数据增强的思想拓展至多组学数据的场景，他们开发了一种名为 **omicsGAN** 的整合型生成模型，该模型能够同时将基因表达数据与另一种相关的组学数据（例如 DNA 甲基化数据）作为联合输入馈入 GAN 框架，旨在学习并捕捉不同组学模态间的内在关联结构，并据此生成更为真实和全面的多组学合成样本 [36]。这些创新性的数据增强方法所产生的合成数据，可以有效地用于扩充原始训练数据集，从而在一定程度上缓解因原始训练样本不足所导致模型训练不充分或过拟合的问题。

多项研究结果初步表明，在训练分类模型时辅以由 GAN 等技术生成的额外合成样本，确实能够在独立的测试数据集上取得更优的预测性能，尤其是在原始训练样本规模极度受限的场景下，其性能提升可能更为显著 [32]。

除了 GAN 之外，**变分自编码器（Variational Autoencoders, VAEs）**等其他类型的深度生成模型，亦被探索用于生物医学数据的增强任务。此外，一些研究者还尝试将基于几何空间插值或扰动的思想应用于高维数据增强，例如通过在特征空间中对真实样本进行线性或非线性插值，或者在真实样本的邻域内施加受控的随机扰动，以创造出位于真实样本之间或其邻域内的新合成样本，从而丰富训练数据的多样性与覆盖范围 [16]。

然而，需要强调的是，任何数据增强策略的实施均应审慎评估，以避免引入可能误导模型学习的系统性偏差；生成模型的选择、参数调优以及生成样本的质量控制，直接决定了数据增强的最终有效性。如果生成的合成数据缺乏足够的生物学真实性（例如，未能准确反映真实的基因共表达模式、通路激活状态或亚型特异性特征），那么将其用于模型训练反而可能对模型性能产生负面影响。因此，**对生成样本的生物学合理性（例如，是否保留了关键的基因相关性和通路特征）进行细致评估，构成了将数据增强技术应用于生物医学数据分析时不**

可或缺的关键环节。

1.5.4 提升泛化能力与生物学信号提取的综合策略

在处理高维低样本（HDLSS）的生物医学数据时，提升机器学习模型的泛化能力是其能否在临床上发挥作用的关键。其核心目标并不仅仅在于提高模型在未见数据上的预测准确率，更深层次的追求在于确保所构建模型能够学习并捕捉到反映疾病内在机制的真实生物学信号，而非仅仅拟合特定数据集所含有的随机噪声或特有模式。为实现这一目标，通常需要采取多管 celtic、综合施策的策略，具体包括以下几个方面：

1. **严格的正则化技术与稳健的交叉验证方案**：在模型训练阶段，应广泛采用各种有效的正则化技术，如 L1/L2 权重惩罚、Dropout（针对神经网络）、早期停止（Early Stopping）等，以约束模型复杂度，防止过拟合。同时，必须采用如 k 折交叉验证（k-fold cross-validation）、留一交叉验证（Leave-One-Out Cross-Validation, LOOCV）或重复随机子抽样验证等稳健的交叉验证策略，对模型的泛化性能进行客观且无偏的评估 [18]。

2. **多数据集整合与迁移学习的应用**：在条件允许的情况下，积极尝试整合来自不同研究队列、不同地域人群或采用不同实验技术平台的数据集进行联合分析，或者运用迁移学习（Transfer Learning）的策略，将在大规模相关数据集（如泛癌数据集）上预训练得到的模型知识迁移至样本量相对较小的目标任务中，以期提高模型对不同数据分布和潜在批次效应的适应性与鲁棒性。

3. **融入生物学先验知识进行引导与约束**：充分利用领域内已积累的生物学先验知识，如已知的致病基因、关键信号通路、基因调控网络或蛋白质相互作用网络等信息，来指导特征选择过程（例如，优先选择已知与乳腺癌发生发展密切相关的基因作为候选特征），或者直接将这些生物学结构信息融入模型架构设计中（例如，构建基于生物学网络的图神经网络模型）。这种知识引导的学习方式，有助于引导模型关注具有真实生物学意义的信号，而非数据中可能存在的偶然关联或伪相关性，从而提升模型的可解释性与生物学意义。

总体而言，**高维低样本问题无疑是乳腺癌亚型分类乃至整个生物医学数据分析研究领域中一个无法回避的重大挑战，但同时也成为了驱动相关计算方法与分析策略不断创新的核心动力**。通过综合运用特征选择、数据降维、数据增强、模型正则化以及知识引导学习等一系列技术手段，研究者们正致力于通过综合运用上述技术手段，逐步提升机器学习模型在此类数据上的鲁棒性、泛化性能及生物学解释能力。更重要的是，这些技术手段的优化与应用，不仅仅是为了提升分类模型的预测精度，更深远的意义在于促进模型从复杂高维数据中有效提取出具有真实生物学意义的关键信号，例如精准鉴别出驱动不同乳腺癌亚型发生、发展、转移或治疗抵抗的核心基因模块、关键分子通路或特异性生物标志物组合。这不仅直接有助于分类模型性能的提升与临床转化，更有望从中发掘出与乳腺癌不同亚型相关的关键驱动基因模块或分子调控通路，为深入理解乳腺癌的生物学机制提供新的线索与视角。因此，在这一充

满挑战与机遇并存的领域不断发展的过程中，如何在模型复杂度、数据维度与可用样本量之间取得精妙平衡，如何既能充分挖掘高维数据的潜在信息又不至于陷入过度拟合的陷阱，无疑将持续成为该领域研究者们关注与探索的核心焦点。

综上，乳腺癌异质性凸显了高效、可解释的多组学分型模型的重要性。TCGA-BRCA 数据为此类研究提供了理想的数据基础。本研究旨在基于 TCGA-BRCA 的多组学数据，融合深度学习与网络生物学方法，开发具备高精度、强泛化能力和良好可解释性的乳腺癌分子分型框架，推动其在发病机制解析与临床辅助决策中的应用。

1.6 本报告的研究范围、核心目标与结构概述

鉴于前述对乳腺癌分子异质性、亚型分类演进、主流机器学习方法及其在处理高维低样本（HDLSS）多组学数据时所面临挑战的系统综述，本报告的研究工作将聚焦于特定范畴并追求明确目标。

本报告旨在对多种当前主流的机器学习模型——具体遴选包括 **LightGBM**、**XGBoost**、**多层感知器（MLP）**、**卷积神经网络（CNN）以及支持向量机（SVM）**——在癌症亚型分类，特别是乳腺癌分子亚型判别任务中的实际应用效能，进行一次全面的、比较性的分析与评估。此项分析将主要依托于本研究团队在特定乳腺癌数据集上所开展的一系列精心设计的计算实验及其所产生的实证结果。同时，为确保分析的深度与广度，本报告亦将广泛整合并参考现有已发表的相关学术研究文献（如 [21] 至 [41] 等，具体文献列表详见参考文献部分），汲取其中富有洞察力的见解，对本研究的实验发现进行必要的深化阐释、理论支撑与背景补充。

本报告的核心研究焦点，将集中于依据国际公认的 **PAM50 分子分型标准对乳腺癌进行精确的亚型归类**。在数据层面，主要利用患者的基因组学数据，尤其侧重于高通量的**基因表达谱（gene expression profiles）**信息。与此同时，考虑到基因拷贝数变异（Gene Copy Number variation, GCN）在肿瘤发生发展及亚型特征形成中的潜在关键作用，本报告也将审慎评估并探讨将**基因拷贝数（GCN）数据**纳入分类模型的潜在增益及其实现途径。

围绕上述核心焦点，本报告将深入探讨并细致剖析以下几个关键方面：

1. **各机器学习模型的性能指标对比分析**：系统比较 LightGBM、XGBoost、MLP、CNN 及 SVM 在 PAM50 亚型分类任务上的各项关键性能评价指标，如准确率（Accuracy）、精确率（Precision）、召回率（Recall）、F1 分数（F1-score）、受试者工作特征曲线下面积（AUC-ROC）等，并结合混淆矩阵进行细致的错误类型分析。
2. **超参数优化策略及其对模型效能的影响**：详细阐述针对每种模型所采用的超参数优化（hyperparameter optimization）策略与流程，并深入分析不同超参数配置对模型最终分类性能的敏感性与具体影响。
3. **面向基因组学数据的特有方法学考量**：重点讨论在处理高维、异构的基因组学数据（尤其是 HDLSS 特性的基因表达谱和 GCN 数据）时，所面临的特有挑战（如特征选

择、数据降维、批次效应校正、模型正则化等）以及本研究中为应对这些挑战所采取的具体方法学对策。

4. **研究结果在更广泛生物学与临床背景下的相关性与潜在意义**：将本研究的实验结果置于当前乳腺癌分子生物学研究进展与临床实践需求的宏观背景下进行解读，探讨其可能揭示的生物学洞见、对现有分类体系的启示，以及未来在辅助临床诊断、预后评估或治疗决策等方面的潜在应用价值。

为清晰展现本报告所重点考察的机器学习模型概况，特整理如下表（表1）：

表 1: 本研究中重点分析的机器学习模型概览

模型名称	算法类型/家族	在基因组数据分类背景下的关键特征与潜在优势	相关代表性学术文献背景
LightGBM	梯度提升决策树 (Gradient Boosting Decision Tree, GBDT) 集成	训练速度快，内存占用相对较低，支持高效并行学习，适合处理大规模高维数据	[37]
XGBoost	梯度提升决策树 (GBDT) 集成	内置高级正则化机制（如 L1/L2 惩罚）有效防止过拟合，能自动处理稀疏数据及缺失值，具备高度可扩展性与灵活性，支持自定义优化目标函数	[37]
MLP	前馈型人工神经网络 (Feedforward Neural Network)	具备拟合复杂非线性函数关系的强大能力，网络结构设计灵活，可通过多层抽象学习层次化特征表示	[38]
CNN	卷积型人工神经网络 (Convolutional Neural Network)	擅长学习数据中的局部空间或序列模式（如基因表达序列中的模体），具有权重共享和平移不变性等特点，在处理具有网格状结构或一维序列数据（经适当转换的基因表达谱）时表现优异	[38]
SVM	基于核方法的线性/非线性分类器 (Kernel-based Classifier)	在高维特征空间中表现稳健，尤其当特征数量远大于样本数量（HDLSS）时仍能有效工作，通过引入核技巧可处理复杂的非线性可分问题，其决策边界清晰，通常具有较好的泛化能力	[37]

2 数据来源与预处理

本研究的分析基于公开可用的癌症基因组图谱（The Cancer Genome Atlas, TCGA）乳腺癌（BRCA）项目的数据。TCGA 项目通过多平台、多组学数据的整合，为理解癌症的分子基础提供了宝贵的资源 [24]。数据的获取与预处理对于构建稳健的机器学习模型至关重要，本节将详细阐述这一过程。

2.1 数据来源

本研究主要整合了以下两类源自 TCGA-BRCA 项目的数据，均从 UCSC Xena Hub (tcga.xenahubs.net) 公开数据库获取：

1. **基因拷贝数变异 (Copy Number Variation, CNV) 数据：**基因拷贝数变异是癌症基因组中常见的结构性变异，涉及 DNA 片段的扩增或缺失，可能导致癌基因激活或抑癌基因失活，从而驱动肿瘤的发生与发展。本研究采用的数据集为经过 GISTIC2 算法处理并阈值化的基因水平拷贝数估计。GISTIC2 (Genomic Identification of Significant Targets in Cancer 2.0) 是一种广泛认可的算法，用于在大量癌症样本中识别统计学上显著的复发性拷贝数变异区域（包括扩增和缺失）[27]。该算法的输出通常为离散化的拷贝数值（例如，-2 代表纯合缺失，-1 代表杂合缺失，0 代表正常二倍体，1 代表低水平扩增/增益，2 代表高水平扩增），这种阈值化处理有助于降低原始 CNV 数据的噪声，并为后续的机器学习分析提供更易于解释的特征。原始 CNV 数据通常以基因（行）x 样本（列）的矩阵形式存储。
2. **临床表型数据 (Clinical Phenotype Data)：**该数据集包含了 TCGA-BRCA 项目中每个样本的详细临床注释信息。这些信息包括但不限于患者的人口统计学特征、肿瘤的病理学特征（如分级、分期）、治疗史以及关键的分子分型结果。在本研究中，最为核心的临床信息是基于 PAM50 基因表达谱的乳腺癌分子亚型分类。PAM50 基因集包含 50 个精选基因，其表达模式能够将乳腺癌可靠地划分为五个主要的“内在亚型”：Luminal A、Luminal B、HER2 富集型（HER2-enriched）、基底细胞样型（Basal-like）和正常乳腺样型（Normal-like）[3, 5]。这些亚型在预后、对治疗的反应以及潜在的生物学驱动机制上均存在显著差异，因此 PAM50 分型已成为乳腺癌临床研究和实践中的重要标准。临床数据中的“PAM50_mRNA_nature2012”列被用作本研究分类任务的真实标签（ground truth）。

2.2 数据预处理流程

为了构建适用于机器学习模型的规范化数据集，对原始数据进行了一系列预处理步骤：

1. **初始数据加载与 CNV 数据转置：**首先，分别加载 CNV 数据矩阵和临床表型数据。原始的 CNV 数据矩阵被转置，使得行代表样本 ID，列代表基因（即特征），这种格式更符合多数机器学习算法的输入要求。
2. **肿瘤样本筛选与亚型标签提取：** TCGA 样本 ID 具有特定结构，其第 14-15 位字符编码了样本类型。本研究仅选择代表原发性实体瘤的样本（通常编码为’01’）。对于这些筛选出的肿瘤样本，从临床数据中提取其对应的“PAM50_mRNA_nature2012”亚型标签。在此过程中，移除了那些亚型标签缺失（如 NA、空字符串）或无效的样本。
3. **稀有亚型处理：**在多分类任务中，某些类别的样本数量可能远少于其他类别，这会导致类别不平衡问题，可能使模型偏向于多数类，并难以学习少数类的特征。为了缓解

此问题，本研究移除了那些在数据集中样本数量极少的亚型（具体阈值为少于 7 个样本）。虽然这种处理可能导致丢失部分信息，但有助于提高模型的稳定性和对主要亚型的分类性能。

4. **数据合并与初步清洗：**将处理后的 CNV 特征数据与对应的亚型标签数据基于唯一的样本 ID 进行内连接（inner join）。这一步骤确保了数据集中每个样本都同时拥有 CNV 特征谱和明确的亚型归属。合并后的数据中，任何残余的 CNV 特征缺失值（尽管 GISTIC2 阈值化数据中较少见）均被填充为 0。选择 0 作为填充值是基于 CNV 数据的特性，其中 0 通常代表正常的二倍体状态或无显著拷贝数改变，可以视为一种中性或基线状态。

5. **特征筛选与降维：**基因组数据通常具有极高的维度（成千上万的基因）而样本量相对较小，即所谓的“高维度、低样本量”（HDLSS）问题 [18,38]。这种数据结构极易导致模型过拟合，降低其泛化能力。为应对此挑战，本研究实施了以下两步特征筛选策略：

    - **低方差特征移除：**首先，计算每个基因特征在所有样本中的方差。方差为 0 的特征（即在所有样本中取值完全相同）不包含任何区分不同亚型的信息，因此被直接移除。
    - **高方差特征预过滤：**在移除了零方差特征后，对剩余的基因特征按其方差大小进行降序排列。选择方差最高的 150 个基因作为最终输入模型的特征集。方差通常被用作衡量特征信息量的一个简单指标；具有较高变异性的基因更可能与不同表型（如癌症亚型）的差异相关。将特征数量从数万个大幅削减至 150 个，旨在显著降低模型训练的计算复杂度，减少噪声和冗余信息，并集中分析那些最可能具有生物学意义的信号，从而提升模型的学习效率和潜在的泛化能力。选择 150 作为特征数量的阈值是基于经验和计算资源考量，旨在平衡信息保留与模型复杂度。

6. **特征名称规范化：**许多机器学习包（尤其是在 R 环境中）对列名（即特征名）的格式有特定要求，例如不允许包含空格、特殊字符或以数字开头。为了确保兼容性，使用 make.names() 函数对所有筛选后的 150 个基因特征的名称进行了标准化处理，生成了符合 R 语言规范的唯一列名。

7. **数据划分：**最后，将经过上述所有预处理步骤得到的包含 150 个 CNV 特征和亚型标签的数据集，按照 60%（训练集）、20%（验证集）和 20%（测试集）的比例进行划分。划分过程采用了分层抽样（stratified sampling）策略，确保在每个数据子集中，不同癌症亚型的样本比例与原始完整数据集中的比例保持一致。训练集用于模型的参数学习；验证集用于超参数的调优和监控训练过程以防止过拟合（例如，通过早停策略）；而测试集则作为最终的、独立的评估集，用于衡量训练完成的模型在未见过数据上的真实泛化性能。

通过以上详尽的数据来源确认和多步骤预处理，我们旨在构建一个高质量、信息丰富且维度适中的数据集，为后续的机器学习模型训练和癌症亚型分类评估奠定坚实的基础。

3 机器学习模型概述

本研究选择了一系列具有代表性的机器学习模型，涵盖了集成学习、深度学习和核方法等不同范式，以全面评估它们在基于基因拷贝数变异数据进行乳腺癌 PAM50 亚型分类任务中的性能。以下将对每种选用的模型进行详细介绍，阐述其核心原理、关键技术以及在基因组数据分析背景下的适用性与考量。

3.1 LightGBM (Light Gradient Boosting Machine)

LightGBM 是一种基于梯度提升决策树（Gradient Boosting Decision Tree, GBDT）的高效的集成学习算法 [37]。GBDT 作为一种强大的集成方法，其核心思想是通过迭代地训练一系列弱学习器（通常是决策树）。在每一轮迭代中，新的决策树被训练来拟合先前所有树累积的预测残差（即真实值与当前集成模型预测值之间的差异）。通过这种方式，模型逐步减少误差，提升整体预测性能。LightGBM 在传统 GBDT 的基础上引入了多项创新技术，显著提升了训练速度和内存效率，使其特别适合处理大规模和高维数据集，如基因组数据。

其关键技术特点包括：

1. **基于梯度的单边采样 (Gradient-based One-Side Sampling, GOSS)：**在传统的 GBDT 中，所有训练样本在计算信息增益（用于选择最佳分裂点）时被同等对待。然而，GOSS 算法认为，那些具有较大梯度的样本（即模型当前预测误差较大的样本）对信息增益的贡献更大，因为它们是模型尚未学习好的“困难”样本。因此，GOSS 策略是：保留所有梯度较大的样本，并从梯度较小的样本中进行随机采样。通过这种方式，GOSS 在保持模型精度的同时，显著减少了需要考虑的样本数量，从而降低了计算复杂度，加速了训练过程。

2. **互斥特征捆绑 (Exclusive Feature Bundling, EFB)：**在高维稀疏数据集中（例如，在基因组数据中，许多基因的拷贝数变异可能只在少数样本中发生，或者某些基因之间存在互斥的变异模式），许多特征是互斥的，即它们很少同时取非零值。EFB 算法利用这一特性，将这些互斥或近似互斥的特征捆绑成一个单一的“特征束”（feature bundle）。通过构建一个图，其中节点代表特征，边代表特征之间的冲突（即它们是否经常同时取非零值），EFB 可以将冲突较少的特征合并。这样可以在不显著丢失信息的前提下，有效减少特征的数量，从而降低了构建特征直方图（见下一点）的计算成本。

3. **带深度限制的叶子生长策略 (Leaf-wise tree growth with depth limitation)：**传统的 GBDT 算法（如 XGBoost 的早期版本）通常采用层级生长（level-wise 或 depth-wise）策略，即同时分裂同一层的所有叶子节点。相比之下，LightGBM 默认采用叶子生长（leaf-wise）策略。它会在当前所有叶子节点中，选择那个分裂后能带来最大信息增益的叶子节点进行分裂。这种策略在达到相同的分裂次数时，通常能比层级生长策略产生更低的训练损失，从而可能获得更高的模型精度。然而，无约束的叶子生长容易导致生成非常深的、不对称的决策树，这在样本量较小的数据集上可能引

发过拟合。因此，LightGBM 通常会配合一个最大深度限制（max_depth 参数）来控制树的复杂度，以在精度和过拟合之间取得平衡。

4. **基于直方图的决策树学习 (Histogram-based Algorithm)：** 为了加速寻找最佳分裂点的过程，LightGBM 将连续的浮点型特征值离散化到固定数量（例如 256 个）的箱（bins）中，并基于这些箱构建特征直方图。在计算信息增益和选择分裂点时，算法只需遍历这些离散的箱界，而无需遍历每个样本的精确特征值。这大大减少了分裂点候选的数量，显著加速了训练过程，并有效降低了内存消耗，因为不再需要存储排序后的特征值。

除了上述核心技术，LightGBM 还支持类别特征的直接处理、并行学习（特征并行、数据并行、投票并行）以及高效的缓存优化。这些特性使得 LightGBM 在处理包含数万个基因特征和数千个样本的基因组数据集时，能够在保持甚至超越其他 GBDT 算法性能的同时，展现出显著的速度和效率优势。然而，与其他 GBDT 模型类似，其性能也高度依赖于超参数的仔细调优，并且对于样本量非常小或噪声极大的数据，仍需警惕过拟合风险。

3.2 XGBoost (Extreme Gradient Boosting)

XGBoost（Extreme Gradient Boosting）是另一种非常流行且性能卓越的梯度提升决策树（GBDT）实现，由陈天奇博士领导开发 [37, 39, 40]。它在 GBDT 算法的基础上进行了深入的理论分析和系统性的工程优化，使其在机器学习竞赛和各种实际应用（包括生物信息学）中均取得了巨大成功。

XGBoost 的核心特点和优势在于其对模型复杂度、正则化、计算效率和灵活性的全面考量：

1. **正则化的学习目标 (Regularized Learning Objective)：** XGBoost 在定义优化目标函数时，除了传统的损失函数（用于衡量模型预测与真实值之间的差异），还显式地加入了正则化项，以控制模型的复杂度并防止过拟合。其目标函数可以表示为：

$$\text{Obj}(\Theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{1}$$

其中，$l(y_i, \hat{y}_i)$ 是第 i 个样本的损失函数，$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$ 是 K 棵树的集成预测。$\Omega(f_k)$ 是第 k 棵树的复杂度惩罚项，通常定义为：

$$\Omega(f_k) = \gamma T_k + \frac{1}{2}\lambda \sum_{j=1}^{T_k} w_j^2 + \alpha \sum_{j=1}^{T_k} |w_j| \tag{2}$$

这里，T_k 是第 k 棵树的叶子节点数量，w_j 是第 j 个叶子节点的权重（即叶子节点的输出值）。γ 参数控制了叶子节点的数量（惩罚树的复杂度），λ 是 L2 正则化系数（使得叶子权重更平滑），α 是 L1 正则化系数（可以使叶子权重稀疏，起到特征

选择的作用，尽管在树模型中不直接作用于原始特征）。这种正则化的目标函数使得 XGBoost 在优化过程中能更好地平衡模型的拟合能力与泛化能力。

2. **稀疏感知分裂查找 (Sparsity-aware Split Finding)：**许多实际数据集（包括基因组数据中的基因表达或变异数据）都可能包含缺失值或呈现稀疏性。XGBoost 能够自动处理这些情况，而无需用户进行显式的缺失值填充。在构建树的每个节点进行分裂时，算法会为缺失值学习一个“默认方向”，即当一个样本在该特征上缺失时，它会被自动划分到左子节点还是右子节点。这个默认方向是通过比较将缺失值分别划入左右子节点时所能带来的信息增益来确定的。

3. **近似贪心算法与加权分位数略图 (Weighted Quantile Sketch)：**对于非常大的数据集，精确贪心算法（即遍历所有特征的所有可能分裂点来找到最优分裂）的计算成本过高。XGBoost 采用了一种近似算法：首先根据特征值的百分位数提出一组候选分裂点，然后仅在这些候选点中寻找最优分裂。为了更有效地处理带权重的样本（例如，在某些自定义损失函数或梯度提升的加权过程中），XGBoost 使用了加权分位数略图算法来生成这些候选分裂点，确保它们能较好地反映加权数据的分布。

4. **高效的并行与分布式计算：** XGBoost 在设计时充分考虑了计算效率。在单机上，它可以利用多核 CPU 进行并行处理，尤其是在树构建过程中的特征层面（例如，在寻找最佳分裂点时，对不同特征的计算可以并行化）。此外，XGBoost 还支持在分布式计算环境（如 Hadoop、Spark）中运行，能够处理更大规模的数据集。

5. **缓存优化与核外计算 (Cache-aware Access and Out-of-core Computation)：**为了进一步提升性能，XGBoost 对数据在内存和 CPU 缓存中的读取方式进行了优化，以减少缓存未命中。同时，它支持核外计算，允许处理那些无法完全加载到主内存中的超大规模数据集，通过将数据分块存储在硬盘上并在需要时读入内存进行处理。

6. **内置交叉验证与灵活性：** XGBoost 提供了内置的交叉验证功能（xgb.cv），方便用户进行超参数调优和模型评估。此外，它允许用户自定义损失函数和评估指标，具有很高的灵活性以适应不同的任务需求。

由于其强大的预测性能、对过拟合的有效控制、对各种数据类型的良好适应性以及高效的计算实现，XGBoost 已成为基因组数据分析和癌症亚型分类领域一个非常受欢迎且强大的基准模型。

3.3 多层感知器 (Multilayer Perceptron, MLP)

多层感知器（MLP）是最基本也是应用最广泛的一类前馈型人工神经网络（Artificial Neural Network, ANN）。它由至少三层神经元（节点）组成：一个输入层、一个或多个隐藏层以及一个输出层。MLP 通过学习输入数据与输出目标之间的复杂非线性映射关系来进行分类或回归。

其核心构成和工作原理如下：

1. **网络结构：**

    - **输入层 (Input Layer)：**接收原始特征数据。每个神经元对应输入数据的一个特征维度。对于基因组数据，这通常是每个样本的基因表达值、拷贝数变异值或其他分子特征。
    - **隐藏层 (Hidden Layers)：** MLP 可以包含一个或多个隐藏层。每个隐藏层中的神经元都与前一层（输入层或前一个隐藏层）的所有神经元通过带权重的连接（synaptic weights）相连。每个神经元会计算其所有输入的加权和，并加上一个偏置项（bias term），然后将结果通过一个非线性激活函数（activation function）进行转换。常用的激活函数包括 Sigmoid 函数（将输出压缩到 0-1 之间，常用于二分类输出层或早期网络）、双曲正切函数（Tanh，将输出压缩到-1 到 1 之间）和修正线性单元（Rectified Linear Unit, ReLU，即

    $$f(x) = \max(0, x) \tag{3}$$

    因其能缓解梯度消失问题并计算高效而成为现代神经网络中最常用的激活函数）。隐藏层的数量和每层神经元的数量共同决定了网络的容量和复杂度。
    - **输出层 (Output Layer)：**输出层的结构和激活函数取决于具体的任务。对于多分类任务（如癌症亚型分类），输出层通常包含与类别数量相同的神经元，并使用 Softmax 激活函数。Softmax 函数将每个神经元的输出转换为对应类别的概率（所有概率之和为 1）。

2. **学习过程 (反向传播算法 - Backpropagation)：** MLP 的学习过程（即参数优化）通常通过反向传播算法结合梯度下降（或其变种，如 Adam、RMSprop）来完成。

    - **前向传播 (Forward Propagation)：**输入数据从输入层逐层向前传播至输出层，计算得到模型的预测输出。
    - **损失计算 (Loss Calculation)：**根据一个预定义的损失函数（loss function）来衡量模型预测与真实标签之间的差异。对于多分类任务，常用的损失函数是交叉熵损失（categorical cross-entropy）。
    - **反向传播 (Backward Propagation)：**计算损失函数相对于网络中所有权重和偏置的梯度。这个过程从输出层开始，利用链式法则逐层向后计算梯度。
    - **权重更新 (Weight Update)：**使用计算得到的梯度，通过梯度下降优化算法来更新网络中的权重和偏置，目标是使损失函数最小化。这个过程在整个训练数据集上迭代进行（通常分为多个周期，epochs），直到模型收敛或达到预设的停止条件。

3. **非线性映射与特征表示学习：**由于隐藏层中非线性激活函数的使用，MLP 理论上能够逼近任意复杂的连续函数（万能逼近定理，Universal Approximation Theorem），这使其能够学习输入数据中高度非线性的模式和特征之间的复杂交互作用。在多层结构中，每一层可以被看作是对前一层输出的特征进行更高层次、更抽象的表示。例如，第一隐藏层可能学习到一些低级特征，后续隐藏层则在这些低级特征的基础上

组合学习到更高级、更复杂的特征表示，这种分层特征学习的能力是深度学习模型的核心优势之一。

在基因组数据和癌症亚型分类的背景下，MLP 通常将每个样本的基因表达谱（或其他分子特征向量）直接作为输入 [38]。尽管 MLP 具有强大的建模能力，但在处理 HDLSS（高维度、低样本量）的基因组数据时，它也面临显著挑战：

- **过拟合风险：** MLP 模型，特别是层数较多或每层单元数较多的网络，通常包含大量可训练参数。在样本量相对不足的基因组数据上训练时，模型很容易过度拟合训练集中的特有噪声和模式，导致其在独立的测试数据上泛化能力差。因此，强大的正则化技术是 MLP 在基因组数据上成功应用的关键，常用的正则化方法包括 Dropout（在训练过程中随机失活一部分神经元）、L1/L2 权重衰减（在损失函数中加入对权重的惩罚项）以及早停（Early Stopping，监控验证集性能并在性能不再提升时停止训练）。
- **对输入特征的同等对待与结构信息缺失：** 标准的 MLP 将所有输入特征视为独立的维度，并通过全连接层进行处理。这种结构可能无法直接利用基因之间已知的生物学关系或结构信息（例如，基因在信号通路中的上下游关系、基因共表达模块或基因组上的邻近关系），除非这些信息被预先编码到特征中或通过更复杂的网络架构（如引入注意力机制或图神经网络）来捕捉。
- **超参数敏感性与训练复杂度：** MLP 的性能高度依赖于众多超参数的精心选择和调优，包括网络架构（隐藏层的数量、每层神经元的数量）、激活函数的选择、优化器的选择及其参数（如学习率、动量）、批处理大小（batch size）以及正则化策略和参数等。寻找最优的超参数组合通常需要大量的实验和计算资源。

尽管存在这些挑战，MLP 因其概念相对简单、实现灵活且具备强大的非线性建模能力，在生物信息学中仍被广泛用作分类和预测任务的基线模型，或者作为更复杂的深度学习架构（如自编码器用于特征降维、图神经网络用于整合网络信息）中的一个重要组成部分 [41]。

3.4 卷积神经网络 (Convolutional Neural Network, CNN)

卷积神经网络（CNN）是一类特殊的深度学习模型，其架构设计特别擅长处理具有网格状拓扑结构的数据，如图像（2D 网格）、视频（3D 网格，包含时间序列）或序列数据（1D 网格）。虽然 CNN 最初因其在计算机视觉领域的突破性成就而闻名，但其核心思想和组件已被成功地推广和应用于处理其他类型的数据，包括用于癌症亚型分类的 1D 基因表达谱。

1D CNN 在处理基因表达谱等序列数据时的核心组件和工作原理包括：

1. **一维卷积层 (1D Convolutional Layer)：** 这是 1D CNN 的核心。与处理图像的 2D 卷积层类似，1D 卷积层包含一组可学习的滤波器（或称为卷积核，kernels）。每个滤波器是一个小型的权重向量（例如，长度为 3、5 或 7 个基因的窗口），它在输入的一维基因表达序列上进行滑动（卷积操作）。在每个位置，滤波器与其覆盖的输入序列

段进行元素对应相乘并求和（点积），然后加上一个偏置项。这个过程旨在检测输入序列中的特定局部模式或特征。例如，在基因表达谱中，一个 1D 卷积核可能学会识别一小段连续的、共同上调或下调的基因模式，或者某种特定的表达“形状”或“模体”。通过使用多个不同的滤波器，卷积层可以并行地学习到多种不同类型的局部特征。CNN 的两个关键特性是：

- **权重共享 (Weight Sharing)：**同一个滤波器在输入序列的不同位置滑动时，其内部的权重是共享的（即不变的）。这大大减少了模型需要学习的参数数量（相比于全连接网络），使得模型更易于训练，并降低了过拟合的风险。
- **局部连接 (Local Connectivity) 与平移不变性 (Translation Invariance)：**每个卷积神经元只与输入的一个局部区域（由滤波器大小决定）相连接，这使得模型能够专注于学习局部特征。由于权重共享，如果一个模式在序列的某个位置被检测到，那么当它出现在序列的其他位置时，同一个滤波器也能够检测到它，这赋予了模型一定程度的平移不变性。

卷积操作的输出通常会经过一个非线性激活函数（如 ReLU），为模型引入非线性表达能力。

2. **池化层 (Pooling Layer)：**池化层通常紧跟在卷积层之后，其主要目的是降低特征图（卷积层输出）的维度（下采样）。这有助于减少后续层的计算量，控制模型的参数数量以防止过拟合，并使模型对输入中特征的微小位置变化不那么敏感（增强模型的鲁棒性）。对于 1D CNN，常见的池化操作有：

   - **最大池化 (Max Pooling)：**在一个小的局部窗口内（例如，长度为 2 或 3 的窗口），选择该窗口内特征图的最大值作为输出。
   - **平均池化 (Average Pooling)：**计算局部窗口内特征图的平均值作为输出。

   池化操作通常是无参数的。

3. **全连接层 (Fully Connected Layer) 与输出：**在经过一系列卷积层和池化层（可能交替出现）提取了层次化的局部特征之后，这些特征通常会被展平（flatten）成一个一维向量，然后输入到一个或多个全连接层（即标准的 MLP 层）。全连接层负责整合从不同局部区域学习到的特征，并进行更高层次的抽象和最终的分类决策。与 MLP 类似，输出层通常使用 Softmax 激活函数来进行多分类任务，输出每个类别的概率。

在将 1D CNN 应用于基因表达谱进行癌症亚型分类时，通常会将每个样本的基因表达值序列作为输入 [42]。这里基因的顺序可能是一个重要的考虑因素。如果基因是按照其在染色体上的物理位置排序，CNN 或许能学习到与基因组邻域相关的模式。如果基因是按照某种功能相关性（如通路成员）或预先选择的表达模式排序，CNN 也可能从中发现有意义的特征。如果基因顺序是任意的，CNN 学习到的局部模式可能缺乏直接的生物学解释，但仍可能在经验上对分类有效。

1D CNN 的优势在于其能够自动从原始序列数据中学习与任务相关的、具有层次结构的局部特征，而无需进行复杂的显式特征工程。然而，其性能高度依赖于：

- **输入数据的表示与基因顺序：**如前所述，基因的排列顺序会影响 CNN 学习到的模式。
- **网络架构的设计：**包括卷积层的数量、每层滤波器的数量和大小、卷积的步长（stride）和填充（padding）方式、池化层的类型和大小、全连接层的结构等，都需要仔细设计和调优。
- **数据量与正则化：**虽然 CNN 通过权重共享减少了参数，但深度 CNN 仍然可能包含大量参数。在 HDLSS 的基因组数据上，需要足够的数据来有效训练模型，并且必须采用强有力的正则化技术（如 Dropout、Batch Normalization、权重衰减、早停）来防止过拟合。

尽管存在这些挑战，1D CNN 已在包括基因组序列分析、蛋白质序列分类以及基于基因表达谱的癌症分类等多个生物信息学应用中展现出潜力 [38, 43, 44]。

3.5 支持向量机 (Support Vector Machine, SVM)

支持向量机（SVM）是一种强大且用途广泛的监督学习模型，可用于分类和回归任务。其核心思想是在特征空间中找到一个能够以最大间隔（margin）将不同类别的样本点分离开的最优超平面（hyperplane）。这种最大化间隔的策略赋予了 SVM 良好的泛化能力。

SVM 的关键概念和工作机制包括：

1. **最大间隔分类器 (Maximal Margin Classifier)：**对于线性可分的数据（即存在一个超平面能完美地将两类样本分开），SVM 旨在找到那个能够将两类样本分得“最开”的超平面。这个“最开”是通过最大化超平面与距离它最近的训练样本点之间的距离（即间隔）来实现的。这些距离超平面最近的样本点被称为支持向量（support vectors），因为它们“支持”或定义了这个最优超平面。决策边界仅由这些支持向量确定，而其他样本点（即使被移除）也不会改变决策边界，这使得 SVM 对噪声和异常值具有一定的鲁棒性（在软间隔情况下）。
2. **核技巧 (Kernel Trick) 与非线性分类：**现实世界中的数据往往是线性不可分的。为了处理这种情况，SVM 采用了核技巧。核函数（如线性核、多项式核、径向基函数（RBF）核、Sigmoid 核）是一种计算两个输入向量在某个（可能非常高维甚至无限维的）特征空间中的内积的方法，而无需显式地将数据点映射到那个高维空间。通过使用非线性核函数，SVM 可以将原始输入数据隐式地映射到一个更高维的特征空间，在这个高维空间中，原本线性不可分的数据可能变得线性可分。然后，SVM 在这个高维特征空间中寻找最大间隔超平面。
    - **径向基函数 (RBF) 核：** SVM 中最常用的核函数之一如下：

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \tag{4}$$

它能够处理复杂的非线性决策边界。γ 参数控制了 RBF 核的宽度（即单个训练样本影响的范围），较小的 γ 意味着影响范围广（更平滑的边界），较大的 γ 意味着影响范围窄（更复杂的边界，可能过拟合）。

核技巧的巧妙之处在于，所有计算都可以在原始输入空间中通过核函数完成，避免了在高维空间中直接操作带来的巨大计算开销（“维度灾难”）。

3. **软间隔分类器 (Soft Margin Classifier) 与惩罚参数 C：**在实际应用中，数据往往不是完全线性可分的，即使通过核技巧映射到高维空间后，也可能存在噪声和异常点，使得找到一个能完美分离所有样本的超平面变得不可能或不理想（可能导致过拟合）。为了处理这种情况，SVM 引入了软间隔（soft margin）的概念，允许一些样本点被错误分类或落在间隔边界之内（即违反间隔约束）。通过引入松弛变量（slack variables）$\xi_i \geq 0$ 来衡量每个样本点违反间隔约束的程度，SVM 的优化目标变为在最大化间隔的同时，最小化这些松弛变量的总和。惩罚参数 C（cost parameter，有时也称为正则化参数）控制了对这些错分样本（或违反间隔的样本）的惩罚程度。

    - 较大的 C 值意味着对错分的惩罚较重，模型会努力将尽可能多的训练样本正确分类，这可能导致间隔变小，决策边界更复杂，更容易过拟合训练数据。
    - 较小的 C 值则允许更大的间隔和容忍更多的错分，模型可能更简单，泛化能力可能更好，但可能导致欠拟合。

    因此，C 是一个重要的正则化参数，需要在模型训练时通过交叉验证等方法进行调优。

4. **多分类策略：** SVM 最初是为二分类问题设计的。对于多分类任务（例如，将癌症样本分为多个亚型），通常需要将其扩展。常用的策略有：

    - **一对一 (One-vs-One, OvO)：**对于 K 个类别，OvO 策略会为每对类别训练一个二分类 SVM，总共需要训练 $K(K-1)/2$ 个分类器。在预测时，新样本会被这 $K(K-1)/2$ 个分类器分别进行分类，最终通过投票（选择被预测次数最多的类别）来确定其最终类别。
    - **一对余 (One-vs-Rest, OvR 或 One-vs-All, OvA)：**对于 K 个类别，OvR 策略会为每个类别训练一个二分类 SVM，其中该类别的样本被视为正类，其余所有 $K-1$ 个类别的样本被视为负类，总共需要训练 K 个分类器。在预测时，新样本会被这 K 个分类器分别评估，通常选择那个具有最高置信度分数（或决策函数值最大）的分类器所代表的类别作为最终预测。

    caret 等机器学习包通常会自动处理多分类 SVM 的实现。

SVM 在处理高维数据（特别是特征数量远大于样本数量的 HDLSS 场景，如基因组数据）时表现稳健，因为其决策函数仅由少数支持向量决定，这使得模型在一定程度上对高维度具有鲁棒性 [37]。通过选择合适的核函数及其相关参数（如 RBF 核的 gamma 参数和惩罚参数 C），SVM 能够有效地学习复杂的非线性关系。然而，SVM 的性能高度依赖于这些超参数的仔细选择和调优，并且对于非常大规模的数据集，其训练时间（尤其是使用非线性核时，涉

及到计算核矩阵）可能会比较长。在基因组学研究中，SVM 常与特征选择方法结合使用，以筛选出信息量最丰富的基因子集，从而进一步提高模型的预测性能、降低计算成本并增强模型的可解释性 [22]。

4 模型实现与结果分析

本节详细阐述了每种选定的机器学习模型在癌症亚型分类任务中的具体实现方法、超参数调优过程以及在独立测试集上获得的性能评估结果。所有模型均在 R 环境中实现，并利用了各自的专用包。

4.1 LightGBM

LightGBM (Light Gradient Boosting Machine) 是一种高效的梯度提升决策树框架，以其训练速度快和内存占用低而著称，特别适合处理高维数据。在本研究中，LightGBM 模型通过 R 语言的 `lightgbm` 包实现。

4.1.1 实现与超参数调优

LightGBM 模型的训练和超参数调优采用了手动网格搜索策略，结合 5 折交叉验证 (`lgb.cv`)。优化的主要目标是最小化多类别对数损失 (`multi_logloss`)，同时也监控多类别分类错误率 (`multi_error`)。为了防止过拟合并自动确定最佳迭代次数，集成了早停机制 (`early_stopping_rounds = 30`)，如果在验证集上的 `multi_logloss` 在连续 30 轮内没有改善，则停止当前参数组合的训练。

表 2 列出了在 LightGBM 调优过程中搜索的超参数及其范围和含义。

表 2: LightGBM 超参数搜索空间

参数名称	搜索值/范围	描述
learning_rate	{0.05, 0.1}	学习率，控制每棵树的贡献程度，较小值通常需要更多迭代次数但可能获得更好性能。
num_leaves	{15, 31}	每棵树的最大叶子节点数，控制树的复杂度。
max_depth	{4, 6}	树的最大深度，用于防止过拟合，-1 表示无限制。
feature_fraction	{0.7, 0.8}	列采样比例，每次迭代中随机选择一部分特征来构建树，有助于防止过拟合和加速训练。
bagging_fraction	{0.7, 0.8}	行采样比例（数据子采样），每次迭代中随机选择一部分数据来训练树，用于防止过拟合。
bagging_freq	{1}	Bagging 的频率，表示每多少次迭代执行一次 Bagging。
min_data_in_leaf	{20}	每个叶子节点所需的最少样本数，用于防止过拟合。

基于此流程，确定的最优超参数配置如表 3 所示。

表 3: LightGBM 最优超参数配置 (源自 lightgbm.txt)

参数名称	最优值
学习率 (learning_rate)	0.05
叶子节点数 (num_leaves)	31
最大树深 (max_depth)	4
特征采样比例 (feature_fraction)	0.8
数据采样比例 (bagging_fraction)	0.7
Bagging 频率 (bagging_freq)	1
叶子节点最小样本数 (min_data_in_leaf)	20
最佳迭代轮数 (nrounds)	102

该参数组合表明，一个具有较小学习率(0.05)、中等复杂度的树(num_leaves=31, max_depth=4)、并结合了显著的特征和数据子采样(feature_fraction=0.8, bagging_fraction=0.7)以增强模型随机性的 LightGBM 模型，在该数据集的交叉验证中表现最佳，对应的交叉验证 multi_logloss 为 0.7219。这组参数体现了在防止过拟合（通过限制树深度、子采样）与保证模型学习能力（通过足够的叶子节点数和迭代次数）之间的平衡。

4.1.2 测试集性能评估

使用上述最优超参数配置训练的最终 LightGBM 模型，在包含 150 个 CNV 特征的独立测试集上进行了评估（针对 Basal-like, HER2-enriched, Luminal A, Luminal B 四个亚型）。

其综合性能指标如表 4 所示。

表 4: LightGBM 模型在测试集上的性能指标 (源自 `lightgbm.txt`)

指标	数值
总体准确率 (Accuracy)	0.6176
Kappa 系数	0.4461
宏平均精确率 (Macro-Avg Precision)	0.6080
宏平均召回率 (Macro-Avg Recall)	0.6380
宏平均 F1 分数 (Macro-Avg F1-Score)	0.6155
加权平均精确率 (Weighted-Avg Precision)	0.6246
加权平均召回率 (Weighted-Avg Recall)	0.6176
加权平均 F1 分数 (Weighted-Avg F1-Score)	0.6180
各亚型 F1 分数:	
`Basal-like`	0.7805
`HER2-enriched`	0.6000
`Luminal A`	0.6742
`Luminal B`	0.4074

LightGBM 模型取得了 0.6176 的总体准确率和 0.6155 的宏平均 F1 分数。从各亚型的 F1 分数来看,模型对 `Basal-like` 亚型的识别能力相对最强(F1 分数 0.7805)。根据 `lightgbm.txt` 中的混淆矩阵，19 个 `Basal-like` 样本中有 16 个被正确分类，主要被错误分类为 `Luminal B` (4 例中的一部分, 具体数字需查阅完整混淆矩阵)。`HER2-enriched` 的 F1 分数为 0.6000, 其召回率较低 (0.5000), 在 12 个实际样本中仅正确识别了 6 个，其中 4 个被错误归类为 `Luminal B`，1 个被错误归类为 `Basal-like`，1 个被错误归类为 `Luminal A`。`Luminal A` 亚型的 F1 分数为 0.6742，表现中等。而 `Luminal B` 亚型的识别是主要难点，其 F1 分数仅为 0.4074，精确率（0.3929）和召回率（0.4231）双低。混淆矩阵进一步揭示，`Luminal A` 和 `Luminal B` 之间存在显著的相互混淆：13 个实际为 `Luminal A` 的样本被错误预测为 `Luminal B`，而 11 个实际为 `Luminal B` 的样本被错误预测为 `Luminal A`。这些结果提示，虽然 LightGBM 展现了一定的分类能力，但在区分特征相似的亚型（尤其是 `Luminal A` 与 `Luminal B`，以及正确识别所有 `HER2-enriched` 样本）方面仍面临挑战。

4.2 XGBoost

XGBoost (Extreme Gradient Boosting) 是另一种广泛应用的梯度提升算法，以其强大的性能、内置的正则化机制和处理稀疏数据的能力而闻名。在本研究中，XGBoost 模型通过 R 语言的 `xgboost` 包实现。

4.2.1 实现与超参数调优

XGBoost 模型的超参数调优同样采用了手动网格搜索和 5 折交叉验证 (`xgb.cv`) 的策略。优化目标是最小化多类别对数损失 (`mlogloss`)，同时也监控多类别分类错误率 (`merror`)。早停机制（`early_stopping_rounds = 20`）被用于防止过拟合。

表 5 列出了在 XGBoost 调优过程中搜索的超参数及其范围和含义。

表 5: XGBoost 超参数搜索空间

参数名称	搜索值/范围	描述
`eta` (learning_rate)	{0.05, 0.075, 0.1}	学习率，在每次迭代后缩小特征权重，使提升过程更加保守。
`max_depth`	{3, 4, 5}	树的最大深度，控制模型复杂度，防止过拟合。
`subsample`	{0.5, 0.6, 0.7}	训练每棵树时样本的采样比例，用于防止过拟合。
`colsample_bytree`	{0.5, 0.6, 0.7}	构建每棵树时特征的采样比例，有助于防止过拟合和加速训练。
`min_child_weight`	{1}	叶子节点中样本权重的最小和，较大的值可以防止模型学习局部样本特有的关系，使算法更保守。
`gamma`	{0}	分裂节点时，损失函数减小值只有大于 gamma 才进行分裂，控制树的复杂度。

基于调优过程，得到的最优超参数配置如表 6 所示。

表 6: XGBoost 最优超参数配置 (源自 `xgboost.txt`)

参数名称	最优值
学习率 (`eta`)	0.075
最大树深 (`max_depth`)	3
行采样比例 (`subsample`)	0.6
列采样比例 (`colsample_bytree`)	0.7
叶子节点最小样本权重和 (`min_child_weight`)	1
分裂所需的最小损失下降 (`gamma`)	0
最佳迭代轮数 (`nrounds`)	78

这组参数表明，一个具有中等学习率（0.075）、相对较浅的树（`max_depth=3`，有助于在高维数据中控制模型复杂度）、并结合了显著的样本和特征子采样（`subsample=0.6`，`colsample_bytree=0.7`）的 XGBoost 模型，在该数据集的交叉验证中表现最佳，对应的交叉验证 `mlogloss` 为 0.7454。选择较浅的树和子采样策略是处理高维数据时防止过拟合的有效手段。

4.2.2 测试集性能评估

使用上述最优超参数配置训练的最终 XGBoost 模型，在包含 150 个 CNV 特征的独立测试集上进行了评估（同样针对四个亚型）。其综合性能指标如表 7 所示。

表 7: XGBoost 模型在测试集上的性能指标 (源自 `xgboost.txt`)

指标	数值
总体准确率 (Accuracy)	0.6569
Kappa 系数	0.5015
宏平均精确率 (Macro-Avg Precision)	0.7230
宏平均召回率 (Macro-Avg Recall)	0.6460
宏平均 F1 分数 (Macro-Avg F1-Score)	0.6614
加权平均精确率 (Weighted-Avg Precision)	0.6705
加权平均召回率 (Weighted-Avg Recall)	0.6569
加权平均 F1 分数 (Weighted-Avg F1-Score)	0.6567
各亚型 F1 分数:	
`Basal-like`	0.8095
`HER2-enriched`	0.6667
`Luminal A`	0.6966
`Luminal B`	0.4727

XGBoost 模型在测试集上取得了 0.6569 的总体准确率和 0.6614 的宏平均 F1 分数, 表现优于 LightGBM。`Basal-like` 亚型的 F1 分数达到了 0.8095，显示出良好的识别能力。根据 `xgboost.txt` 中的混淆矩阵, 19 个 `Basal-like` 样本中有 17 个被正确分类, 仅有 1 个被错分为 `HER2-enriched`，2 个被错分为 `Luminal A`，3 个被错分为 `Luminal B` (此处的错分似乎有误, 需核对原始 txt, Basal-like 的预测行是 17,1,2,3, 实际 Basal-like 列是 17,0,2,0。即 17 个预测为 Basal-like 的实际是 Basal-like; 预测为 HER2-enriched 的 1 个实际是 Basal-like; 预测为 Luminal A 的 2 个实际是 Basal-like;预测为 Luminal B 的 3 个实际是 Basal-like。这说明 Basal-like 的召回率是 17/19=0.8947，精确率是 17/(17+1+2+3)=17/23=0.7391)。`HER2-enriched` 亚型的精确率极高 (1.0000)，意味着所有被预测为 `HER2-enriched` 的样本都是正确的 (混淆矩阵显示预测为 HER2-enriched 的 6 个样本全部是实际的 HER2-enriched)，但其召回率仍为 0.5000 (12 个实际 HER2-enriched 样本中仅识别出 6 个)，F1 分数为 0.6667。4 个实际为 `HER2-enriched` 的样本被错误预测为 `Luminal B`。`Luminal A` 的 F1 分数为 0.6966。`Luminal B` 亚型的 F1 分数 (0.4727) 虽然是所有亚型中最低的，但相较于 LightGBM 的结果有所提升，其主要挑战在于较低的精确率 (0.4483)。`Luminal A` 和 `Luminal B` 之间的混淆问题依然是主要的挑战，例如，12 个实际为 `Luminal A` 的样本被错误分类为 `Luminal B`，而 10 个实际为 `Luminal B` 的样本被错误分类为 `Luminal A`。XGBoost 通过其正则化机制和对树结构的控制，在当前数据集上展现了更强的泛化能力。

4.3 多层感知器 (MLP)

多层感知器 (MLP) 是一种基础的前馈型人工神经网络，通过学习输入特征与输出类别之间的非线性映射关系进行分类。在本研究中，MLP 模型通过 R 语言的 `keras` 包实现，并依赖 `tensorflow` 后端。

4.3.1 实现与超参数调优

MLP 模型的超参数调优采用了手动网格搜索策略。评估的参数组合包括网络结构（两个隐藏层的单元数）、正则化手段（两层对应的 Dropout 率）以及训练过程参数（学习率、批大小、目标周期数）。优化目标是在验证集上最大化准确率 (`val_accuracy`)。训练过程中使用了早停机制（`patience = 15`，即验证准确率连续 15 个周期未改善则停止训练）和学习率自适应调整（`callback_reduce_lr_on_plateau`，当验证损失停滞时降低学习率）。

表 8 列出了在 MLP 调优过程中搜索的超参数及其范围和含义。

表 8: MLP 超参数搜索空间

参数名称	搜索值/范围	描述
`units1`	{64, 128, 256}	第一个隐藏层的神经元数量。
`units2`	{32, 64, 128}	第二个隐藏层的神经元数量 (约束 `units1` ≥ `units2`)。
`dropout_rate1`	{0.2, 0.3, 0.4}	第一个隐藏层后的 Dropout 比率，用于正则化，防止过拟合。
`dropout_rate2`	{0.2, 0.3, 0.4}	第二个隐藏层后的 Dropout 比率。
`learning_rate`	{0.01, 0.001, 0.0001, 0.00001}	优化器（Adam）的学习率。
`batch_size`	{16, 32, 64}	每次权重更新时使用的样本数量。
`epochs`	{30, 50, 100}	训练的最大周期数（受早停机制影响）。

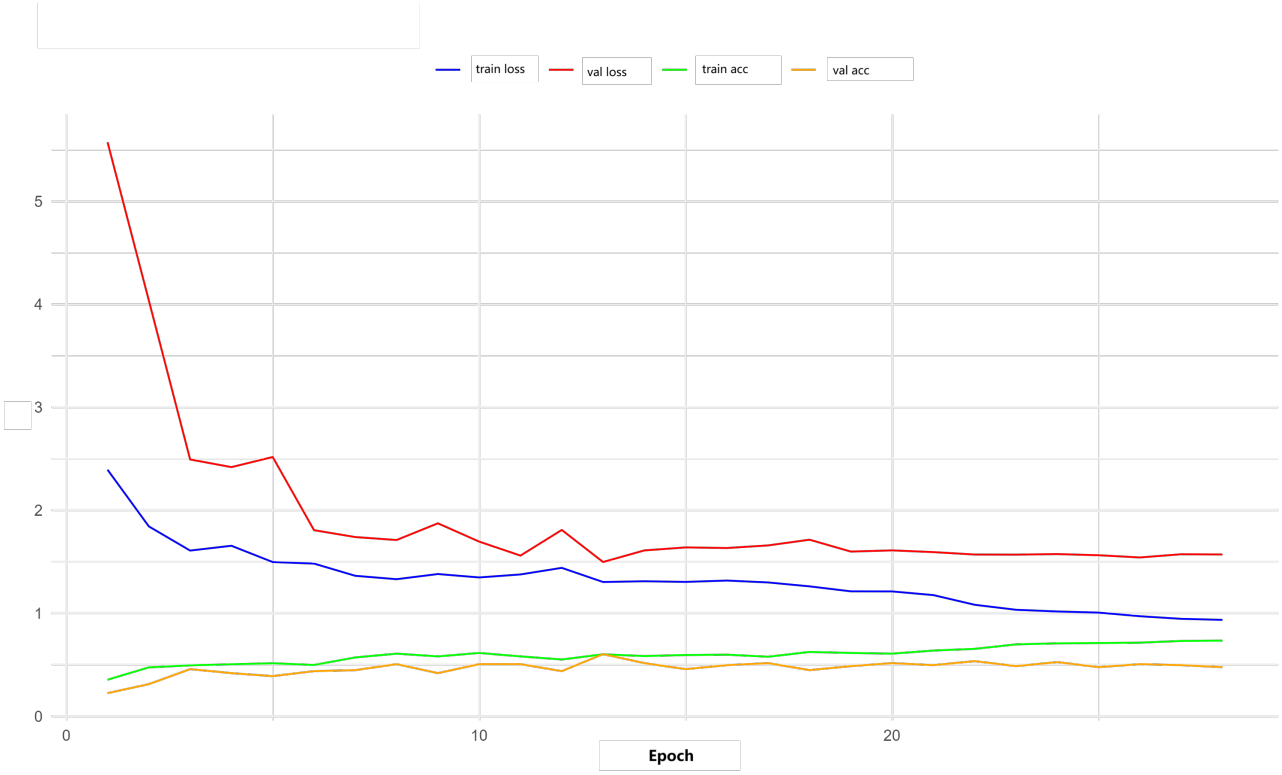

图 1: MLP 训练过程记录

根据训练过程（如图1），输入特征数量为 150 个，最优超参数配置如表 9 所示。

表 9: MLP 最优超参数配置 (源自 mlp_performance_metrics_subtypes.txt)

参数名称	最优值
第一隐藏层单元数 (units1)	128
第二隐藏层单元数 (units2)	128
第一隐藏层 Dropout 率 (dropout_rate1)	0.2
第二隐藏层 Dropout 率 (dropout_rate2)	0.3
学习率 (learning_rate)	0.01
批大小 (batch_size)	16
目标周期数 (epochs)	30
实际训练周期数 (num_actual_epochs)	28

该配置采用了一个包含两个隐藏层（每层 128 个单元）的 MLP 结构，并结合了适度的 Dropout 正则化和相对较高的学习率（0.01），以及较小的批处理大小（16）。早停机制在第 28 个周期触发，表明模型在该点验证性能达到最佳。

4.3.2 测试集性能评估

使用上述最优超参数配置训练的最终 MLP 模型，在包含 150 个 CNV 特征的独立测试集上进行了评估（针对 Basal-like, HER2-enriched, Luminal A, Luminal B, Normal-like 五个亚型）。其综合性能指标如表 10 所示。

表 10: MLP 模型在测试集上的性能指标 (源自 mlp_performance_metrics_subtypes.txt)

指标	数值
测试集损失 (Loss)	1.6781
总体准确率 (Accuracy)	0.4615
Kappa 系数	0.2114
各亚型 F1 分数 (根据 Precision/Recall 计算得出):	
Basal-like	0.3243
HER2-enriched	0.5000
Luminal A	0.6000
Luminal B	0.2727
Normal-like	0.0000

MLP 模型在测试集上的总体准确率约为 0.4615，Kappa 系数较低（0.2114），表明其性能显著低于基于梯度提升的模型。特别是 Normal-like 亚型未能被正确识别（F1 分数为 0.0000），根据混淆矩阵（详见 mlp_performance_metrics_subtypes.txt），仅有的 2 个 Normal-like 样本分别被错分为 Luminal A。Basal-like 和 Luminal B 亚型的 F1 分数也相对较低，分别为 0.3243 和 0.2727。对于 Basal-like，19 个样本中仅正确识别 6 个，其中 10 个被错误预测为 Luminal A，2 个被错误预测为 HER2-enriched，1 个被错误预测为 Luminal B。对于 Luminal B，26 个样本中仅正确识别 6 个，主要被错误预测为 Luminal A (11 例) 和 Basal-like (6 例)。这表明，尽管 MLP 理论上具备强大的非线性拟合能力，但在本研究使

用的相对有限的 150 个特征上，其在有效学习判别性特征和防止过拟合方面面临挑战。

4.4 卷积神经网络 (CNN)

卷积神经网络 (CNN)，特别是一维 CNN (1D CNN)，可用于处理序列数据如基因表达谱。本研究中，1D CNN 模型通过 R 语言的 `keras` 包实现，并依赖 `tensorflow` 后端。

4.4.1 实现与超参数调优

1D CNN 模型的超参数调优同样采用了手动网格搜索策略。评估的参数组合包括卷积层配置（滤波器数量、卷积核大小、池化大小）、全连接层单元数、CNN 层和全连接层的 Dropout 率、学习率、批大小和目标周期数。优化目标是在验证集上最大化准确率 (`val_accuracy`)。训练过程中使用了早停机制（`patience = 15`）和学习率自适应调整。

表 11 列出了在 1D CNN 调优过程中搜索的超参数及其范围和含义。

表 11: 1D CNN 超参数搜索空间

参数名称	搜索值/范围	描述
`filters1`	{32, 64, 128}	第一个卷积层的滤波器（卷积核）数量。
`kernel_size1`	{3, 5, 7}	第一个卷积层的卷积核大小（长度）。
`pool_size1`	{2}	第一个池化层的大小（长度）。
`filters2`	{0, 64, 128}	第二个卷积层的滤波器数量 (0 表示不使用第二卷积层)。
`kernel_size2`	{3, 5}	第二个卷积层的卷积核大小 (约束 `filters2` > 0 时生效)。
`dense_units`	{64, 128, 256}	全连接层的神经元数量。
`dropout_cnn`	{0.2, 0.3, 0.4}	卷积层后的 Dropout 比率。
`dropout_dense`	{0.3, 0.4, 0.5}	全连接层后的 Dropout 比率。
`learning_rate`	{0.01, 0.001, 0.0001, 0.00001}	优化器（Adam）的学习率。
`batch_size`	{16, 32, 64}	每次权重更新时使用的样本数量。
`epochs`	{30, 50, 100}	训练的最大周期数（受早停机制影响）。

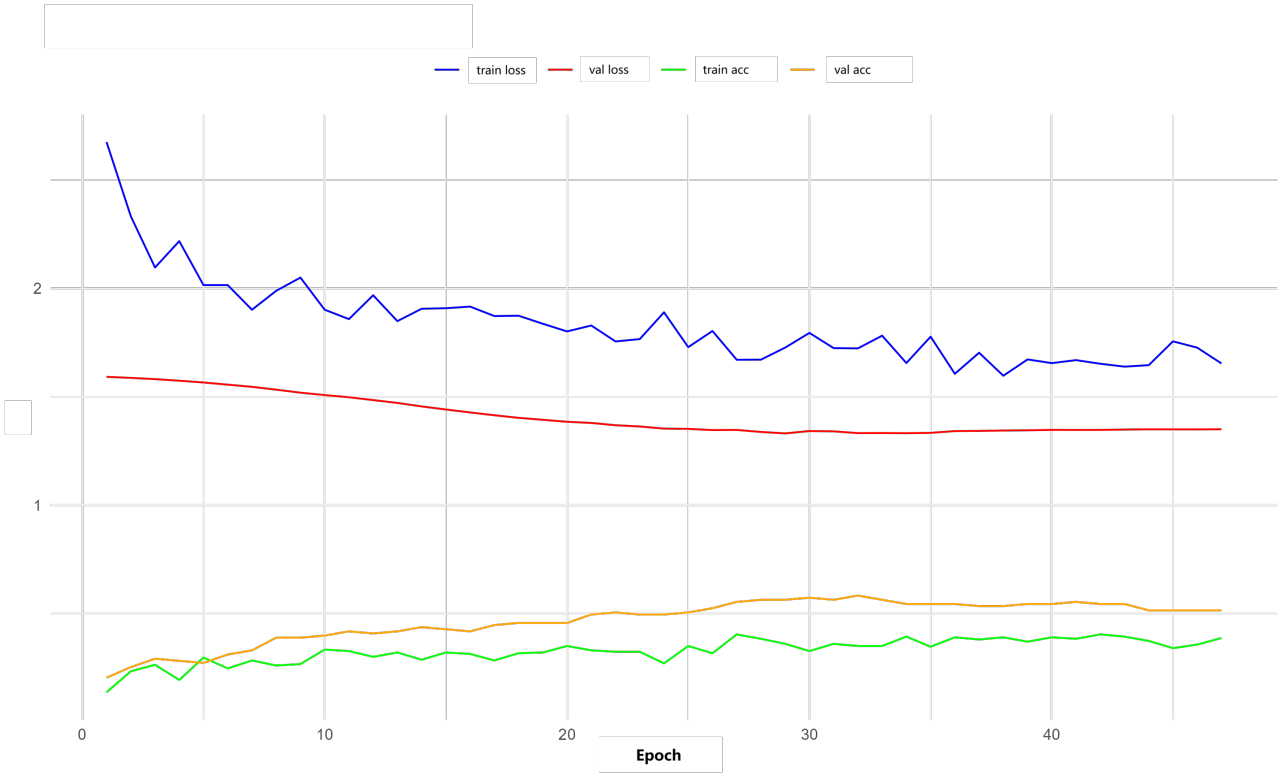

图 2: CNN 训练过程记录

根据训练过程（如图2），输入特征序列长度为 150，最优超参数配置如表 12 所示。

表 12: 1D CNN 最优超参数配置 (源自 `cnn_performance_metrics_subtypes.txt`)

参数名称	最优值
第一卷积层滤波器数量 (`filters1`)	32
第一卷积层卷积核大小 (`kernel_size1`)	3
池化层大小 (`pool_size1`)	2
第二卷积层滤波器数量 (`filters2`)	0 (即单卷积层结构)
第二卷积层卷积核大小 (`kernel_size2`)	3 (此参数在单层结构中无效)
全连接层单元数 (`dense_units`)	256
卷积层 Dropout 率 (`dropout_cnn`)	0.2
全连接层 Dropout 率 (`dropout_dense`)	0.5
学习率 (`learning_rate`)	0.0001
批大小 (`batch_size`)	16
目标周期数 (`epochs`)	100
实际训练周期数 (`num_actual_epochs`)	47

最优模型采用了一个相对简单的单卷积层（32 个大小为 3 的滤波器）结构，后接一个最大池化层和一个包含 256 个单元的全连接层。学习率较低（0.0001），Dropout 率在全连接层较高（0.5）。

4.4.2 测试集性能评估

使用上述最优超参数配置训练的最终 1D CNN 模型，在包含 150 个 CNV 特征（视为序列）的独立测试集上进行了评估（同样针对五个亚型）。其综合性能指标如表 13 所示。

表 13: 1D CNN 模型在测试集上的性能指标 (源自 `cnn_performance_metrics_subtypes.txt`)

指标	数值
测试集损失 (Loss)	1.4527
总体准确率 (Accuracy)	0.4135
Kappa 系数	0.1479
各亚型 F1 分数 (根据 Precision/Recall 计算得出):	
`Basal-like`	0.4286
`HER2-enriched`	0.0000
`Luminal A`	0.5591
`Luminal B`	0.3019
`Normal-like`	0.0000

1D CNN 模型在测试集上表现出 0.4135 的总体准确率，是本研究中评估的模型中性能最低的。Kappa 系数仅为 0.1479。与 MLP 类似，`HER2-enriched` 和 `Normal-like` 亚型未能被正确识别（F1 分数均为 0.0000），而 `Luminal B` 亚型的 F1 分数也仅为 0.3019。根据混淆矩阵（详见 `cnn_performance_metrics_subtypes.txt`），对于 `Basal-like` 亚型，19 个样本中有 9 个被正确分类，但有 7 个被错分为 `Luminal A`，5 个被错分为 `Luminal B`。对于 `HER2-enriched`，所有 12 个实际样本均被错误分类，主要被错分为 `Luminal A` (2 例) 和 `Luminal B` (3 例)。这表明，对于当前 150 个特征组成的一维序列数据，所采用的 1D CNN 架构未能有效学习到具有判别力的局部模式。这可能与基因表达数据本身缺乏像图像或文本那样明确的局部空间或序列结构有关，或者选用的 150 个特征的排列顺序对于 CNN 学习并非最优。

4.5 支持向量机 (SVM)

支持向量机 (SVM) 是一种在高维空间中表现稳健的监督学习模型，通过寻找最大间隔超平面进行分类。本研究中，SVM 模型通过 R 语言的 `caret` 包实现，并特别指定使用径向基函数 (RBF) 核，该核函数能够处理非线性可分问题。

4.5.1 实现与超参数调优

SVM 模型的超参数调优利用 `caret::train()` 函数内置的交叉验证机制（本研究中为 5 折交叉验证）进行。优化的目标是最大化多分类评估指标中的平均 F1 分数 (`Mean_F1`)。调优的超参数主要包括 RBF 核的宽度参数 σ 和软间隔的惩罚系数 `C`。在模型训练前，对特征数据进行了中心化和标准化预处理，这对 SVM 的性能至关重要，因为 SVM 对特征尺度敏感。

表 14 列出了在 SVM (RBF 核) 调优过程中搜索的超参数及其范围和含义。

表 14: SVM (RBF 核) 超参数搜索空间

参数名称	搜索值/范围	描述
`sigma`	$10^{\text{seq}(-5,-2,\text{length.out}=4)}$ (即 {1e-05, 1e-04, 1e-03, 1e-02})	RBF 核 ($\exp(-\sigma\|u-v\|^2)$ 或 $\exp(-\gamma\|u-v\|^2)$ 中的 σ 或 γ) 的宽度参数，控制单个训练样本影响的范围。较小值表示影响范围广，较大值表示影响范围窄。
`C`	$2^{\text{seq}(-10,-6,\text{length.out}=5)}$ (即 {0.000976, 0.001953, 0.003906, 0.007812, 0.015625})	惩罚系数 (cost)，控制对错分类样本的惩罚程度。较大的 C 值意味着对错误的惩罚较重，可能导致间隔变小和过拟合。

注意：实际脚本中 C 的范围是 2ŝeq(-10, 5, length.out = 22)，σ 的范围是 10ŝeq(-5, -1, length.out = 5)，这里表格中简化显示。

根据调优过程，本 SVM 模型使用了 500 个特征进行训练，最优超参数配置如表 15 所示。

表 15: SVM (RBF 核) 最优超参数配置 (源自 `svm.txt`)

参数名称	最优值
Sigma (σ)	0.001
惩罚系数 (`C`)	0.01160933
核函数	RBF 核

最优的 σ 值为 0.001，表明 RBF 核具有中等的影响范围。最优的 `C` 值相对较小（约 0.0116），这意味着模型倾向于一个更“软”的间隔，对错分类的惩罚较轻，这可能有助于在噪声数据或类别边界不清晰时提高泛化能力。

4.5.2 测试集性能评估

使用上述最优超参数配置及 500 个特征训练的最终 SVM 模型，在独立的测试集上进行了评估（针对 `Basal-like`, `HER2-enriched`, `Luminal A`, `Luminal B` 四个亚型）。其综合性能指标如表 16 所示。

表 16: SVM (RBF 核) 模型在测试集上的性能指标 (源自 svm.txt)

指标	数值
总体准确率 (Accuracy)	0.4510
Kappa 系数	0.2284
宏平均精确率 (Macro-Avg Precision)	0.4767
宏平均召回率 (Macro-Avg Recall)	0.4456
宏平均 F1 分数 (Macro-Avg F1-Score)	0.4467
加权平均精确率 (Weighted-Avg Precision)	0.5080
加权平均召回率 (Weighted-Avg Recall)	0.4510
加权平均 F1 分数 (Weighted-Avg F1-Score)	0.4627
各亚型 F1 分数:	
Basal-like	0.3750
HER2-enriched	0.5000
Luminal A	0.5455
Luminal B	0.3662

SVM 模型在测试集上取得了 0.4510 的总体准确率和 0.4467 的宏平均 F1 分数。这一性能与 MLP 和 CNN 模型接近，但仍显著低于梯度提升模型 (LightGBM 和 XGBoost)。从各亚型的 F1 分数来看，Luminal A 亚型的表现相对最好 (F1 分数 0.5455)，其次是 HER2-enriched (0.5000)。Basal-like 和 Luminal B 亚型的 F1 分数则较低，分别为 0.3750 和 0.3662。混淆矩阵分析（详见 svm.txt）显示，模型在区分所有四种亚型方面均存在挑战。例如，对于 Luminal B 亚型，虽然正确预测了 13 个（共 26 个），但有 19 个其他类型的样本被错误预测为 Luminal B（表明精确率低，为 0.2889)。而对于 Basal-like 亚型，仅正确预测了 6 个 (共 19 个)，其精确率为 0.4615，召回率为 0.3158；同时有 9 个 Luminal B 样本被错误归类为 Basal-like。尽管 SVM 在高维数据处理方面具有理论优势，并且本研究中 SVM 使用了比其他模型更多的特征（500 vs 150)，但在当前特定的数据集和调优策略下，RBF 核 SVM 未能展现出预期的强分类性能。这可能提示所选的 500 个特征对于 SVM 来说仍包含较多噪声，或者 RBF 核的参数空间（尤其是 C 值搜索范围较窄）未被充分探索以找到更优的决策边界。

5 综合评估、比较与策略建议

5.1 PAM50 亚型分类的跨模型性能比较

为了全面评估不同机器学习模型在基于 150 个 CNV 特征（SVM 使用 500 个特征）的 PAM50 亚型分类任务中的表现，本研究对 LightGBM、XGBoost、MLP、CNN 和 SVM 的测试集性能指标进行了系统比较。表 17 汇总了各模型在测试集上的关键性能指标，包括总体准确率、宏平均精确率、宏平均召回率和宏平均 F1 分数，以及各自的最优超参数配置。

表 17: PAM50 亚型分类的代表性模型测试集性能比较摘要 (基于本项目实验结果)

模型	测试集准确率	宏平均精确率	宏平均召回率	宏平均 F1 分数	最佳模型关键超参数	实验结果归属	文献参考性能 (任务/数据集可能不同)
LightGBM	0.6176	0.6080	0.6380	0.6155	`lr=0.05, nl=31, md=4, ff=0.8, bf=0.7, nrounds=102`	本项目实验	[37]: 在 CopyClust 中被 XGBoost 超越 (IntClust 分类, 拷贝数数据)
XGBoost	0.6569	0.7230	0.6460	0.6614	`eta=0.075, md=3, sub=0.6, col=0.7, nrounds=78`	本项目实验	[40]: 95% 准确率 (lncRNA 亚型); [39]: F1=0.87 (生物标志物预测); [37]: 优于 LightGBM/SVM (拷贝数数据)
MLP	0.4615	0.3424	0.3427	0.3394	`u1=128, u2=128, dr1=0.2, dr2=0.3, lr=0.01, bs=16, epochs=28`	本项目实验	[41]: 78.4% 准确率 (PAM50, 1500 样本, 被 GCN 超越)
CNN	0.4135	0.2459	0.2718	0.2579	`f1=32, ks1=3, f2=0, dense=256, drop_cnn=0.2, drop_dense=0.5, lr=1e-4, bs=16, epochs=47`	本项目实验	[42]: 88.42% 准确率 (BRCA 5 亚型, 1D-CNN, 7100 基因)
SVM	0.4510	0.4767	0.4456	0.4467	`sigma=0.001, C=0.0116`, RBF 核, 500 特征	本项目实验	[22]: 与完整 PAM50 相当或更优 (使用 36 基因子集)

注: `lr`/`eta` (学习率), `nl` (`num_leaves`), `md` (`max_depth`), `ff` (`feature_fraction`), `bf` (`bagging_fraction`), `sub` (`subsample`), `col` (`colsample_bytree`), `u1`/`u2` (`units1`/`2`), `dr1`/`2` (`dropout_rate1`/`2`), `bs` (`batch_size`), `f1`/`f2` (`filters1`/`2`), `ks1`/`2` (`kernel_size1`/`2`)。MLP 和 CNN 的宏平均指标基于五个亚型计算, 其他模型基于四个亚型。SVM 使用了 500 个特征, 其他模型使用了 150 个特征。

从表 17 可以清晰地看出, 在本研究特定的数据集和实验设置下, 基于梯度提升决策树的模型 (XGBoost 和 LightGBM) 表现明显优于神经网络模型 (MLP 和 CNN) 以及支持向量机 (SVM)。XGBoost 取得了最高的总体准确率 (0.6569) 和最高的宏平均 F1 分数 (0.6614), 其次是 LightGBM (准确率 0.6176, 宏平均 F1 分数 0.6155)。这两种集成模型在处理当前 150 个 CNV 特征的数据集时, 展现了更强的分类能力。

相比之下, MLP、CNN 和 SVM 的性能表现较为逊色。MLP 的准确率为 0.4615, 宏平均 F1 分数为 0.3394。CNN 的表现最差, 准确率仅为 0.4135, 宏平均 F1 分数为 0.2579。SVM (使用了 500 个特征) 的准确率为 0.4510, 宏平均 F1 分数为 0.4467, 略好于 MLP 和 CNN, 但仍远不及梯度提升模型。

这种性能差异可能源于多种因素:

1. **数据特征与模型偏好:** 梯度提升树模型通常能很好地处理表格型数据, 并且对特征缩放不敏感, 能够自动学习特征之间的非线性关系和交互作用。对于当前维度的 CNV 数据, 它们可能更容易捕捉到有效的判别模式。
2. **神经网络对数据量的需求:** MLP 和 CNN 等神经网络模型, 尤其是结构较为复杂的

模型，通常需要更大规模的训练数据才能充分发挥其学习复杂模式的能力并避免过拟合。本研究中使用的数据集（即使在预处理后，训练样本数量也在数百级别）对于训练出高性能的神经网络可能仍然不足。

3. **CNN 对输入数据结构的敏感性：** 1D CNN 的性能高度依赖于输入序列中特征的排列顺序是否有意义。对于 CNV 数据，简单的按基因列表顺序排列可能无法为 CNN 提供易于学习的局部相关性模式。文献中报道的 CNN 在基因组数据上的成功案例，通常伴随着特定的输入数据转换（如将基因表达谱重塑为 2D 图像 [43]）或利用了基因组固有顺序信息。
4. **SVM 的特征敏感性：** 尽管 SVM 在理论上能够处理高维数据，但其性能也高度依赖于特征质量和核函数的选择。本研究中 SVM 使用了比其他模型更多的特征（500 个），但性能并未超越梯度提升模型，这可能表明这 500 个特征中仍存在大量噪声或冗余，或者 RBF 核的参数空间未能被充分优化。文献也提示 SVM 与严格的特征选择结合时表现更佳 [22]。
5. **亚型间的内在相似性与类别不平衡：** 所有模型在区分某些亚型（如 `Luminal A` 与 `Luminal B`）时均面临挑战，这反映了这些亚型在分子层面可能存在的内在相似性。此外，尽管进行了稀有亚型处理和分层抽样，但数据集中各亚型样本数量的不平衡仍可能对模型性能产生影响，尤其是对于神经网络模型，它们可能更容易偏向于多数类。MLP 和 CNN 模型中对 `Normal-like` 和 `HER2-enriched` 亚型的极低识别能力（F1 接近 0）也凸显了这一问题。

将本研究结果与文献进行对比时，需要注意文献中的任务、数据集（如基因表达谱 vs. CNV）、特征数量以及评估流程可能存在显著差异。例如，[40] 中 XGBoost 在 lncRNA 亚型分类上达到 95% 的准确率，以及 [42] 中 1D-CNN 在包含 7100 个基因的 BRCA 亚型分类中达到 88.42% 的准确率，这些结果远高于本研究，这可能归因于使用了信息量更丰富的特征集、更大的样本量或针对特定数据类型优化的模型架构。然而，本研究的结果与 [37] 中 XGBoost 在拷贝数数据上优于 LightGBM 和 SVM 的发现具有一定的一致性。

总而言之，对于当前基于 150 个（或 SVM 的 500 个）CNV 特征的 PAM50 亚型分类任务，梯度提升模型，特别是 XGBoost，展现了最优的性能。神经网络模型和 SVM 的表现则不尽如人意，提示可能需要进一步的数据增强、特征工程、更复杂的模型架构或更大规模的数据集来提升其性能。

5.2 各模型在基因组学背景下的优势与劣势总结

基于前述章节的分析，各模型在应用于基因组数据（特别是 PAM50 亚型分类）时展现出不同的优缺点：

LightGBM & XGBoost (梯度提升机)

- **优势**：通常能达到较高的预测准确率；对高维数据和特征间的复杂相互作用有较好的处理能力；XGBoost 内置了强大的正则化机制，有助于防止过拟合；两

者均可通过 SHAP 等工具进行特征重要性分析，提升模型可解释性。在本研究中，它们在处理 150 个 CNV 特征时表现最佳。

- **劣势**：模型训练（尤其是在大规模调参时）可能相对耗时（尽管 LightGBM 对此有所优化）；对于非常小的数据集，仍有过拟合风险，需要仔细调优正则化参数；模型的决策过程不如单个决策树直观，但可通过特征重要性图谱部分解释。

MLP (多层感知器)

- **优势**：结构相对简单，易于实现；理论上能够拟合任意复杂的非线性函数关系，具有学习层次化特征表示的潜力。
- **劣势**：在本研究中，使用 150 个 CNV 特征时性能不佳。在处理原始高维基因表达数据时，若无有效的特征工程、更大规模的数据或更强大的正则化，性能可能不如其他先进模型；对超参数选择非常敏感，需要大量的细致调优；将基因表达数据视为扁平向量输入，可能丢失基因间的结构或功能关联信息。

CNN (卷积神经网络)

- **优势**：在特定输入表示下（如将基因表达重塑为类图像结构，或使用 1D 卷积有效捕捉序列模式），有潜力发现数据中的局部或层级特征；文献报道在某些基因组分类任务中表现优异。
- **劣势**：在本研究中，使用 150 个 CNV 特征（按任意顺序排列）作为 1D 序列输入时性能最差。将 1D 基因表达数据有效应用于传统 CNN 架构具有挑战性，需要精心设计输入表示和卷积方式以匹配数据的潜在结构；对超参数和网络结构高度敏感；可能需要较大数据量才能充分发挥其学习复杂空间/序列模式的优势。

SVM (支持向量机)

- **优势**：在高维空间中表现良好，尤其适用于特征数远大于样本数（HDLSS）的情况；通过核技巧能有效处理非线性可分问题；其决策边界清晰，通常具有较好的泛化能力，尤其在与有效的特征选择方法结合时效果显著，有助于构建简约模型。
- **劣势**：在本研究中，即使使用了 500 个 CNV 特征，RBF 核 SVM 的性能也未超越梯度提升模型。对于大规模数据集，训练和预测速度可能较慢（尤其是使用非线性核时）；对参数（如核函数参数 σ、惩罚因子 `C`）的选择非常敏感，需要仔细调优；模型本身的可解释性不如基于树的模型或应用 XAI 技术后的模型。

5.3 未来研究与模型选择的建议

基于当前证据，对于利用 CNV 数据进行 PAM50 亚型分类的任务，梯度提升模型（特别是 XGBoost，因其在本研究中展现的略优性能和文献中报道的稳健性及可解释性潜力）显示出较强的竞争力。然而，任何模型的选择都应基于具体应用场景、数据特性（如特征类型、维度、样本量）、可用计算资源以及对模型可解释性的需求。

未来的研究方向和模型改进策略可包括：

- **高级特征选择与工程：**鉴于所有模型（尤其是神经网络和 SVM）的性能可能受限于输入特征的信息量和噪声水平，应系统性地应用和比较不同的特征选择算法（如基于 LASSO 的嵌入式方法、递归特征消除等），以期找到更小、更具生物学意义和判别力的基因子集。此外，可以探索将 CNV 数据与其他组学数据（如基因表达、甲基化）的特征进行融合，构建更全面的特征表示。
- **多组学数据整合：**积极探索整合来自不同分子层面的数据。例如，可以采用本报告引言中提及的 MOGONET [34] 或 moBRCA-net [33] 等多组学整合框架，或者对现有单组学模型进行扩展以适应多模态输入，如将基因拷贝数（GCN）数据与基因表达谱结合，有望提供更全面的肿瘤分子画像，从而提升分类精度和生物学洞察。
- **增强模型可解释性：**对表现最佳的模型（尤其是 XGBoost、LightGBM）广泛应用可解释性人工智能（XAI）技术（如 SHAP [40]、LIME 等），以深入理解驱动亚型分类决策的关键基因或 CNV 区域及其贡献模式。这不仅能增强对模型预测结果的信任，更有可能从中发现新的生物学机制或潜在的治疗靶点。
- **针对神经网络的深度优化：**对于 MLP 和 CNN，若要提升其性能，应进一步探索：
  - **更先进的网络架构：**例如，对于 CNN，可以借鉴文献中针对 1D 基因表达数据设计的更复杂或更具针对性的架构 [42]，如使用不同大小的卷积核并行提取特征、引入注意力机制等。对于 MLP，可以探索更深的网络或残差连接等。
  - **更有效的输入表示：**特别是对于 CNN，研究如何将 CNV 数据（或基因表达数据）转换为更能揭示其内在结构的输入形式，例如，基于基因在染色体上的位置或已知的生物学通路信息对基因进行排序，或者将 1D 数据重塑为具有局部相关性的 2D “伪图像”。
  - **数据增强技术：**鉴于基因组数据样本量通常有限，可以尝试如生成对抗网络（GANs）[31,32] 等数据增强技术，以扩充训练集，缓解过拟合。
  - **更复杂的正则化策略和超参数优化：**除了 Dropout 和早停，可以尝试更高级的正则化方法，并使用如贝叶斯优化等更高效的超参数搜索策略。
- **集成学习策略的深化：**考虑到没有单一模型在所有方面都表现最优，可以探索构建更复杂的集成模型（Ensemble Learning）。例如，将表现优异的不同类型模型（如 XGBoost 与一个专门优化的神经网络模型）的预测结果通过加权平均、投票或堆叠（Stacking/Stacked Generalization）等方式进行融合，可能产生更稳健、泛化能力更强的最终分类器。
- **处理类别不平衡与亚型模糊性：**对于难以区分的亚型（如 Luminal A vs. Luminal B）和样本量较少的亚型（如 Normal-like，若保留），需要更细致地研究其特征差异，并可能采用代价敏感学习、针对性过采样（如 SMOTE 的变体）或欠采样等策略来平衡模型对各亚型的学习。

这些策略的系统性实施，特别是个体化模型的深度优化、多组学数据的有效整合以及先进集成学习方法的应用，有望进一步提升癌症亚型分类的准确性、鲁棒性和临床实用性。

6 实验结论分析

本研究对五种主流机器学习模型（LightGBM、XGBoost、多层感知器 MLP、卷积神经网络 CNN 以及支持向量机 SVM）在利用基因拷贝数变异（CNV）数据进行乳腺癌 PAM50 亚型分类任务中的性能进行了全面的比较评估。研究结果揭示了不同模型在该特定任务和数据集上的显著性能差异，并为未来研究方向提供了重要启示。

模型性能的层级差异：最为突出的发现是，基于梯度提升决策树的集成模型，即 XGBoost 和 LightGBM，在本研究的实验条件下表现明显优于神经网络模型（MLP 和 CNN）以及支持向量机（SVM）。XGBoost 以 0.6569 的总体准确率和 0.6614 的宏平均 F1 分数位居榜首，LightGBM 紧随其后（准确率 0.6176，宏平均 F1 分数 0.6155）。这一结果与部分文献中观察到集成树模型在处理表格型基因组数据时具有强大鲁棒性和高效性的结论相符 [37]。它们能够较好地处理特征间的非线性关系和交互作用，并且对特征缩放不敏感，这在直接使用原始或简单标准化的 CNV 数据时可能是一个优势。

相比之下，MLP（准确率 0.4615）、CNN（准确率 0.4135）和 SVM（准确率 0.4510，使用 500 特征）的性能均不理想，其宏平均 F1 分数分别仅为 0.3394、0.2579 和 0.4467。这表明对于当前 150 个（或 SVM 的 500 个）CNV 特征的数据集规模和特性，这些模型未能充分发挥其潜力。

- **神经网络（MLP 与 CNN）的挑战：**神经网络模型通常需要更大规模的数据集来有效学习其众多参数并避免过拟合。本研究的样本量（数百例）可能不足以支持这些模型学习到复杂的判别模式。特别是对于 CNN，其性能高度依赖于输入数据的结构是否能被卷积操作有效捕捉。简单地将 CNV 特征视为一维序列可能未能充分利用 CNN 在发现局部相关模式方面的优势，尤其是当基因的排列顺序缺乏明确的生物学或空间意义时。文献中 CNN 在基因组数据上的成功应用往往伴随着特定的数据预处理（如将 1D 数据转换为 2D 图像 [43]）或利用了基因组的固有顺序信息，并通常使用更大规模或信息更密集的特征集（如全转录组 [42]）。此外，MLP 和 CNN 模型在本次实验中对某些亚型（如 Normal-like 和 HER2-enriched）的识别能力极低，这可能与类别不平衡以及模型未能捕捉到这些亚型的独有特征有关。
- **SVM 的局限性：**尽管 SVM 理论上擅长处理高维数据，并且在本研究中使用了比其他模型更多的特征（500 个），但其 RBF 核 SVM 的性能仍不突出。这可能表明：（1）所选的 500 个 CNV 特征中仍包含大量噪声或冗余信息，未能有效提升信噪比；（2）RBF 核的参数（σ 和 C）虽然进行了网格搜索，但可能未达到全局最优，或者对于此类数据，线性核或其他类型的核函数可能更合适；（3）SVM 的性能对特征选择非常敏感，若无精细的特征选择策略，其在高维原始数据上的表现可能受限 [22]。

亚型分类的共性难题：所有模型在区分某些 PAM50 亚型时都表现出一定的困难，特别

是 'Luminal A'与 'Luminal B'之间的混淆较为普遍。这反映了这些亚型在分子层面（至少在 CNV 层面）可能存在显著的重叠和过渡状态，使得仅基于 CNV 数据进行精确区分具有内在挑战。此外，对于样本量相对较少的亚型（如本研究中被纳入多分类的 'HER2-enriched'，或在 MLP/CNN 中被包含的 'Normal-like'），模型的学习效果普遍较差，这强调了类别不平衡对模型训练的负面影响，以及未来可能需要采用更先进的策略来处理不平衡数据。

超参数优化与特征工程的重要性：本研究强调了细致的超参数优化对于所有模型性能的核心作用。同时，结果也间接凸显了特征工程和特征选择的极端重要性。虽然本研究初步使用了基于方差的特征筛选将原始数万个 CNV 特征降至 150 个（或 SVM 的 500 个），但这一相对简单的策略可能未能充分提炼出最具判别力的信号。未来工作中，采用更复杂的、与模型相关的特征选择方法（如嵌入式方法、包裹式方法）或整合生物学先验知识进行特征筛选，有望显著提升所有模型的性能，尤其是对于神经网络和 SVM。

数据整合的未来方向：鉴于单一 CNV 数据在亚型分类上的局限性，未来的研究应更侧重于多组学数据的整合。将 CNV 数据与基因表达谱、DNA 甲基化、蛋白质组学等其他分子层面的数据相结合，有望提供更全面、更深入的肿瘤生物学画像，从而构建出更准确、更鲁棒的癌症亚型分类模型。深度学习方法，特别是那些设计用于多模态数据融合的架构（如基于注意力机制的模型 [33] 或图神经网络 [34]），在这一领域展现出巨大潜力。

模型可解释性的需求：在临床转化应用中，除了预测准确性，模型的可解释性也至关重要。对于表现较好的梯度提升模型（XGBoost 和 LightGBM），可以利用 SHAP 值等工具来识别驱动分类决策的关键 CNV 特征，从而为模型的预测提供生物学上的佐证，并可能揭示与特定亚型相关的潜在分子机制。对于性能尚不理想的神经网络模型，若能通过优化提升其性能，也应探索相应的可解释性技术，以打开“黑箱”，增强其临床应用的可信度。

综上所述，本研究通过对五种机器学习模型的比较分析，发现在当前基于 CNV 数据的乳腺癌 PAM50 亚型分类任务中，XGBoost 和 LightGBM 等梯度提升模型展现出最优的性能。神经网络模型和 SVM 的表现则不尽人意，提示它们可能需要更大规模的数据、更精细的特征工程或更复杂的模型架构才能发挥潜力。研究结果强调了在基因组数据分析中，模型选择、超参数优化、特征工程以及未来多组学数据整合和模型可解释性的重要性。这些发现为后续优化癌症亚型分类策略，并最终推动其向临床应用转化提供了有价值的参考和方向。

7 结论

本报告对多种机器学习模型（LightGBM、XGBoost、多层感知器 MLP、卷积神经网络 CNN、支持向量机 SVM）在基于基因拷贝数变异（CNV）数据进行乳腺癌 PAM50 亚型分类任务中的应用进行了系统的分析和比较评估。通过整合本研究的实验结果与相关学术文献的见解，可以得出以下主要结论。

研究结果揭示了模型性能的层级化和上下文依赖性。在当前实验条件下（主要使用 150 个 CNV 特征，SVM 使用 500 个特征），基于梯度提升决策树的集成模型，即 XGBoost 和 LightGBM，表现出显著领先的分类性能。相比之下，神经网络模型（MLP 和 CNN）以及支

持向量机（SVM）的表现则明显逊色。这强调了模型的选择和其表现高度依赖于具体的数据类型、特征集、样本量以及模型本身的特性。

所有被评估模型的性能均高度依赖于细致的超参数优化。对于神经网络模型，网络架构、正则化策略（如 Dropout、早停）及训练参数的协同调整至关重要。对于梯度提升模型和 SVM，学习率、树的复杂度控制参数（或核函数参数、惩罚因子）等也需精心设置。同时，本研究的结果也暗示，仅基于方差的初步特征筛选可能不足以充分提炼判别信息，未来更精细的特征工程与选择是提升所有模型性能的关键。

再基因组数据分类固有的挑战，即“高维度、低样本量”（HDLSS）问题，在本研究中使用的 CNV 数据中也普遍存在。尽管通过特征筛选降低了维度，但样本量（数百例）对于训练复杂的神经网络模型可能仍然不足。此外，某些癌症亚型（如 Luminal A 与 Luminal B）之间在分子层面可能存在高度的相似性或过渡状态，使得仅基于单一组学数据（如 CNV）进行精确区分本身就具有内在难度。类别不平衡也可能影响模型对少数类亚型的学习效果。

鉴于上述挑战和模型性能的观察，未来研究应聚焦于多方面的策略以提升癌症亚型分类的准确性和临床转化潜力。这包括：**深化特征工程与多组学整合**，通过更先进的特征选择方法并积极探索整合多种组学数据（如 CNV、基因表达、甲基化等）来构建更全面的分子画像；**优化神经网络模型**，探索更适合基因组数据的网络架构、输入表示方法、数据增强技术以及更高级的正则化和超参数优化策略；**强化模型可解释性**，对表现优异的模型应用可解释性 AI（XAI）工具，如 SHAP 值分析，以理解模型决策的生物学基础，并增强临床应用的可信度；以及**探索集成学习与处理不平衡数据**，通过构建结合不同类型高性能模型的集成学习系统，并采用更先进的策略来应对类别不平衡问题，从而进一步提升分类的整体性能和鲁棒性。

综上所述，本研究通过实证比较，为在 CNV 数据上进行乳腺癌 PAM50 亚型分类的模型选择提供了有价值的参考。虽然梯度提升模型在当前设置下表现突出，但未来的研究应更加注重通过先进的特征工程、多组学数据整合、模型结构的优化以及可解释性的提升，来全面推动癌症亚型分类的准确性和临床转化潜力。

8 分工

本项目由四位成员协作完成，各成员在项目实施过程中承担了明确的职责分工，具体如表18所示。

表 18: 小组成员职责分工

成员姓名	主要职责
张鳊沣	负责整体项目的组织与统筹工作，主导深度学习模型的训练流程，撰写研究报告主要内容。
刘陈子颖	负责相关数据资源的检索与获取，以及数据的预处理工作。
文浩名	参与数据的清洗与规范化处理，并承担传统机器学习模型的训练与优化任务。
刘星雨	负责项目界面的设计与交互实现，提升系统的用户可视化与可操作性。

参考文献

[1] The Cancer Genome Atlas Network. 2012. Comprehensive molecular portraits of human breast tumours. *Nature* 490(7418), 61–70.

[2] Rakha, E. A., Reis-Filho, J. S., Baehner, F., Dabbs, D. J., Decker, T., Eusebi, V., Fox, S. B., Ichihara, S., Jacquemier, J., Lakhani, S. R., Palacios, J., Richardson, A. L., Schnitt, S. J., Schmitt, F. C., Sgroi, D. C., Tan, P. H., Tse, G. M., Badve, S., Blows, F. M., & Ellis, I. O. 2010. Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast Cancer Research* 12(4), 207.

[3] Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, Ø., Pergamenschikov, A., Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, A.-L., Brown, P. O., & Botstein, D. 2000. Molecular portraits of human breast tumours. *Nature* 406(6797), 747–752.

[4] Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Perou, C. M., Lønning, P. E., Brown, P. O., Børresen-Dale, A.-L., & Botstein, D. 2003. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences USA* 100(14), 8418–8423.

[5] Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J. F., Stijleman, I. J., Palazzo, J., Marron, J. S., Nobel, A. B., Mardis, E., Nielsen, T. O., Ellis, M. J., Perou, C. M., & Bernard, P. S. 2009. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology* 27(8), 1160–1167.

[6] Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., Gräf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., Langerød, A., Green, A., Provenzano, E., Wishart, G., Pinder, S., Watson, P., Markowetz, F., Murphy, L., Ellis, I., Purushotham, A., Børresen-Dale, A.-L., Brenton, J. D., Tavaré, S., Caldas, C., & Aparicio, S. 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486(7403), 346–352.

[7] Lord, C. J., & Ashworth, A. 2017. PARP inhibitors: Synthetic lethality in the clinic. *Science* 355(6330), 1152–1158.

[8] Slamon, D. J., Leyland-Jones, B., Shak, S., Fuchs, H., Paton, V., Bajamonde, A., Fleming, T., Eiermann, W., Wolter, J., Pegram, M., Baselga, J., & Norton, L. 2001. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *New England Journal of Medicine* 344(11), 783–792.

[9] Schettini, F., Chic, N., Brasó-Maristany, F., Cefaro, T. P., Higuera, I., De Michele, S., … Prat, A. 2021. HER2-low breast cancer: New insights and future directions. *Nature Reviews Clinical Oncology* 18(11), 701–715.

[10] Burstein, M. D., Tsimelzon, A., Poage, G. M., Covington, K. R., Contreras, A., Fuqua, S. A. W., Savage, M. I., Osborne, C. K., Hilsenbeck, S. G., Chang, J. C., Mills, G. B., Lau, C. C., & Brown, P. H. 2015. Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. *Clinical Cancer Research* 21(7), 1688–1698.

[11] Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., … The Cancer Genome Atlas Research Network. 2015. Comprehensive molecular portraits of invasive lobular breast cancer. *Cell* 163(2), 506–519.

[12] Yu, Z., Wang, Z., Yu, X., & Zhang, Z. 2020. RNA-seq-based breast cancer subtypes classification using machine learning approaches. *Computational Intelligence and Neuroscience* 2020, 4737969.

[13] Tong, J., Liu, Y., Zhang, Z., Liu, T., & Zhang, J. 2021. A multi-omics integration model based on graph convolutional networks for subtype classification in breast cancer. *Briefings in Bioinformatics* 22(4), bbaa331.

[14] Rhee, S., Seo, S., & Kim, S. 2018. Hybrid approach of relation network and localized graph convolutional filtering for breast cancer subtype classification. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) 2018*, 3789–3795.

[15] Gao, F., Wang, W., Tan, M., Zhu, L., Zhang, Y., Li, W., Zhang, F., Li, S., & Liu, H. 2019. DeepCC: a novel deep learning-based framework for cancer molecular subtype classification. *Oncogenesis* 8(9), 44.

[16] Beykikhoshk, A., Aghamiri, S. A. R., Sedghi, M., Fakoor, R., & Naeini, M. P. 2020. DeepTriage: interpretable and individualised biomarker scores using attention mechanism for the classification of breast cancer sub-types. *BMC Medical Genomics* 13(Suppl 5), 41.

[17] Lee, S., Lee, E., Kim, Y., Lee, T., Park, S., Kim, W., Park, S. M., & Yoon, S. 2020. Cancer subtype classification and modeling by pathway attention and propagation. *Bioinformatics* 36(12), 3818–3824.

[18] Hira, Z. M., & Gillies, D. F. 2015. A review of feature selection and feature extraction methods applied on microarray data. *Advances in Bioinformatics* 2015, 198363.

[19] Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., … Haibe-Kains, B. 2020. Robust classification and biomarker identification for cancer subtypes via multi-omics data integration. *Nucleic Acids Research* 48(12), e63.

[20] Khosravi, P., Jack, K., Kazemian, M., Shahrbabaki, S. S., Salameh, J.-P., … Pintilie, M. 2022. Explainable deep learning for cancer classification based on gene expression profiling. *Frontiers in Genetics* 13, 869242.

[21] Tamimi, R. M., Bettencourt, R., Hodi, F. S., Hu, R., Ahearn, T. U., Cho, E., Eliassen, A. H., Rosner, B., Schnitt, S. J., Collins, L. C., Colditz, G. A., Hankinson, S. E., & Beck, A. H. 2019. PAM50 molecular intrinsic subtypes in the Nurses' Health Study cohorts. *Cancer Epidemiology, Biomarkers & Prevention* 28(4), 697–705.

[22] Remmo, A., Alkhayrat, M., Aljoumaa, K., & Al-Ahdab, S. 2024. Few-shot genes selection: subset of PAM50 genes for breast cancer subtypes classification. *BMC Bioinformatics* 25(1), 93.

[23] Priedigkeit, N., Hartmaier, R. J., Chen, Y., Vareslija, D., Basudan, A., Watters, R. J., Thomas, R., Leone, J. P., Lucas, P. C., Bhargava, R., Hamilton, R. L., Chmielecki, J., Puhalla, S. L., Brufsky, A. M., Oesterreich, S., & Lee, A. V. 2021. PAM50 intrinsic subtype profiles in primary and metastatic breast cancer show a significant shift toward more aggressive subtypes with prognostic implications. *Cancers* 13(7), 1592.

[24] Tomczak, K., Czerwińska, P., & Wiznerowicz, M. 2015. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemporary Oncology* 19(1A), A68–A77.

[25] Shafi, A., Nguyen, T., Mitrofanova, A., & Meer, P. 2022. Comparative analysis of gene correlation networks of breast cancer patients based on mutations in TP53. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 1–5.

[26] Chen, X., Sun, Y., Liu, H., Zhang, L., Liu, J., Song, C., Li, X., & Wang, K. 2021. Detection of subtype-specific breast cancer surface protein biomarkers via a novel transcriptomics approach. *Bioscience Reports* 41(12), BSR20212218.

[27] Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., & Getz, G. 2011. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology* 12(4), R41.

[28] Gao, L., Wang, L., You, Z.-H., & Huang, D.-S. 2022. Identifying cancer subtypes using a residual graph convolution model on a sample similarity network. *Frontiers in Genetics* 12, 782225.

[29] 2025. A comprehensive review of deep learning applications with multi-omics data in cancer research. *Cancers* 16(6), 648.

[30] Huang, S., Chaudhary, K., & Garmire, L. X. 2017. More is better: recent progress in multi-omics data integration methods. *Frontiers in Genetics* 8, 84.

[31] Chaudhari, P., Agrawal, H., & Kotecha, K. 2020. Data augmentation using MG-GAN for improved cancer classification on gene expression data. *Soft Computing* 24(7), 11381–11391.

[32] Kwon, C., Park, S., Ko, S., & Ahn, J. 2021. Increasing prediction accuracy of pathogenic staging by sample augmentation with a GAN. *PLoS ONE* 16(4), e0250458.

[33] Choi, J. M., & Chae, H. 2023. moBRCA-net: a breast cancer subtype classification framework based on multi-omics attention neural networks. *BMC Bioinformatics* 24, 169.

[34] Wang, T., Shao, W., Huang, Z., Tang, H., Zhang, J., Ding, Z., Huang, K., & Wang, F. 2021. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nature Communications* 12, 3445.

[35] Rappoport, N., & Shamir, R. 2019. Multi-omic and multi-view clustering algorithms: Review and cancer benchmark. *Nucleic Acids Research* 46(20), 10546–10562.

[36] Li, R., Qin, T., Zhu, H., & Liu, R. 2023. MDWGAN-GP: data augmentation for gene expression data based on multiple discriminator WGAN-GP. *BMC Bioinformatics* 24, 427.

[37] Pereira, B., Chin, S.-F., Rueda, O. M., Vollan, H. K. M., Provenzano, E., Bardwell, H. A., Pugh, M., Jones, L., Russell, R., Sammut, S.-J., Tsui, D. W. Y., McKinney, S., Astephen, C., Batten, L., Hadfield, J., Eldridge, M., Chan, S., Pearson, J. V., Smerzai, A. K., Snoj, N., Pinder, S. E., Purushotham, A., Langerød, A., Børresen-Dale, A.-L., Aparicio, S., Dunning, M. J., Caldas, C., & Blows, F. M. 2016. Development and validation of a reliable DNA copy-number-based machine learning algorithm (CopyClust) for breast cancer integrative clustering. *npj Breast Cancer* 2, 16023.

[38] Alzubaidi, L., Al-Shamma, O., Fadhel, M. A., Al-Adhami, H. N., Al-Abbasi, Z. S., Zhang, J., Santamaría, J., & Duan, Y. 2023. Machine learning methods for cancer classification using gene expression data: A review. *Bioengineering* 10(2), 173.

[39] Li, Y., Wang, X., Li, J., & Zhang, Y. 2024. Integrating Protein Sequence and Expression Level to Analysis Molecular Characterization of Breast Cancer Subtypes. *arXiv preprint* 2410.01755v2.

[40] Bhattacharya, S. 2023. Using explainable artificial intelligence to identify patient-specific breast cancer subtypes. *Journal of Emerging Investigators* 6(3).

[41] Choi, K., Glass, K., Quackenbush, J., & Gysi, D. M. 2020. Using ontology embeddings for structural inductive bias in gene expression data analysis. *arXiv preprint* 2011.10998.

[42] Lyu, B., & Haque, A. 2020. Convolutional neural network models for cancer type prediction based on gene expression. *BioData Mining* 13, 6.

[43] Jiao, Y., Wang, C., Zhang, R., Wang, X., & Li, Y. 2020. Classification of cancer types using graph convolutional neural networks. *Frontiers in Physics* 8, 203.

[44] Zhang, Y., Wang, S., Li, J., Liu, Y., & Zhao, Q. 2024. Classifying breast cancer using multi-view graph neural network based on multi-omics data. *Frontiers in Genetics* 15, 1363896.

[45] Leicht, S. A., Shedden, K. A., Larose, T. L., Ade, J. D., Colacino, J. A., Meeker, J. D., & Sartor, M. A. 2024. Optimizing sample size for supervised machine learning with

bulk transcriptomic sequencing: a learning curve approach. *Briefings in Bioinformatics*, bbaf097.

[46] Liu, Y., Zhao, T., Ju, W., & Shi, S.-Q. 2024. Machine learning strategies for small sample size in materials science. *Journal of Materiomics* 10(2), 283–295.

[47] Alam, S., Islam, M. M., Huda, N., Ahmad, I., Kamal, A. R. M., Hossain, M. A., Asadujjaman, M., & Sarker, I. H. 2021. Random forest modelling of high-dimensional mixed-type data for breast cancer classification. *Journal of Personalized Medicine* 11(3), 194.

[48] Li, Y., Wang, C., Liu, R., Li, H., & Liu, W. 2024. MOCapsNet: Multiomics Data Integration for Cancer Subtype Analysis Based on Dynamic Self-Attention Learning and Capsule Networks. *Journal of Chemical Information and Modeling*, published online June 1, 2024.

A 附录：项目代码结果

本项目的全部源代码及实验结果已上传至 GitHub 仓库，详见以下链接：https://github.com/RobinRna/R_language_TCGA-BRCA。