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 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
	文浩名	中国地质大学
	刘星雨	北京体育大学

汇报人：张鲴沣 | 汇报时间：2025/06/22

目录

01

PART ONE

01 研究背景与问题提出

02 数据获取与方法应用

03 模型实现与综合评估

04 实验总结与结论思考

1.1 研究背景与问题提出

乳腺癌（Breast Cancer），女性常见的恶性肿瘤，发自乳腺上皮细胞的增殖失控。癌症高发因素有遗传以及雌二醇暴露，例如早初潮、晚绝经、不孕以及儿童期胸部受放射线照射等。除了高发人群，晚发育、哺乳期长、早期生育有降低乳癌的机会。肿瘤类型分非浸润癌及浸润癌、且能发展至全身许多器官。发病年龄从20岁起逐年升高，45-50岁达最高点，全球女性杀手第一名，中国乳腺癌病例占全世界的30%，近年呈逐年增长趋势。

乳腺癌其内在特征呈现出**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]。

乳腺癌其生物学行为、疾病进程和治疗反应存在显著的个体差异

精准医疗时代对乳腺癌进行精确分子分型是实现个体化治疗的关键

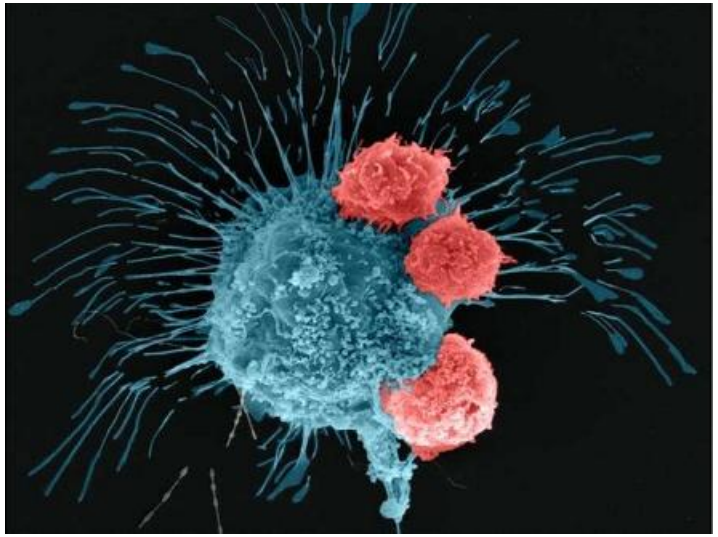

图1.1 扫描电子显微镜图像显示乳腺癌细胞（青色）收到T细胞（红色）攻击

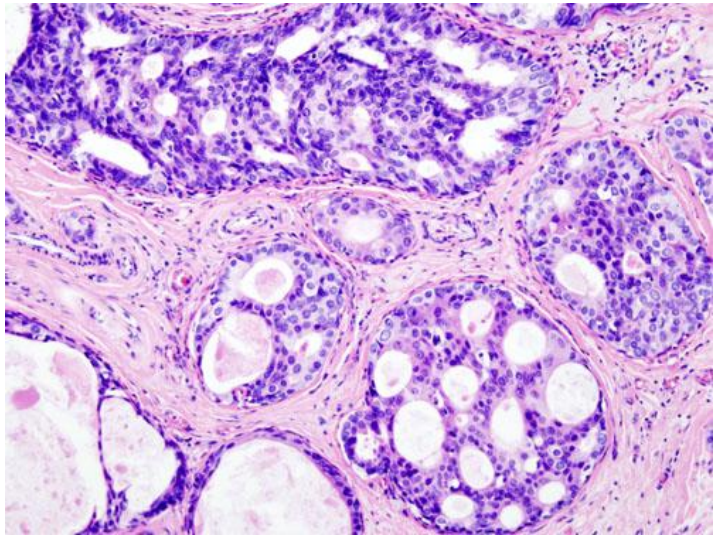

图1.2 导管细胞癌原位（英语：ductal cell carcinoma in situ, DCIS）组织病理图像，苏木精-曙红染色

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

1.1 研究背景与问题提出

乳腺癌传统分型方法[2] （基于免疫组织化学 IHC）

表1 乳腺癌传统分型方法亚型分类

临床亚型	ER/PR 表达	HER2 表达	Ki-67	特点
Luminal A	+	−	低	激素敏感、预后良好
Luminal B	+	±	高	激素敏感、增殖活跃、预后较差
HER2阳性型	−	+	高	易转移、可靶向HER2治疗
三阴性（TNBC）型	−	−	高	侵袭性强、缺乏靶向治疗手段

- 雌激素受体（ER）
- 孕激素受体（PR）
- Ki-67 增殖指数
- 人类表皮生长因子受体2（HER2）

优点：方法简单、易操作，成本低，适合临床广泛使用。

局限：依赖主观判读，难以反映肿瘤的分子异质性；某些病例难以归类。

乳腺癌 PAM50 分型方法[4] （基于基因表达谱）

表2 乳腺癌 PAM50 分型方法亚型分类

PAM50 亚型	主要特征	预后
Luminal A	激素受体阳性、增殖低、基因表达稳定	最佳
Luminal B	激素受体阳性、增殖高、部分HER2表达	中等
HER2-enriched	HER2通路激活但可能HER2蛋白阴性	较差
Basal-like	类似TNBC，细胞分裂活跃、p53突变多	最差
Normal-like	接近正常乳腺组织，临床意义不明确	待定

PAM50 分型是一种基于50个关键乳腺癌相关基因的表达模式对肿瘤进行分类的分子分型方法[3-4]，代表了“内在亚型（intrinsic subtype）”的分型标准。

优点：基于客观的高通量基因表达谱，分类更精细稳定；可以揭示同一IHC亚型中隐藏的分子异质性；治疗反应预测能力强。

局限：成本较高；需要特定平台和标准化流程。

PAM50分型通过基因表达谱精细揭示乳腺癌内在异质性，是实现个体化精准医疗的重要基础

[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.

1.2 方法背景与问题挑战

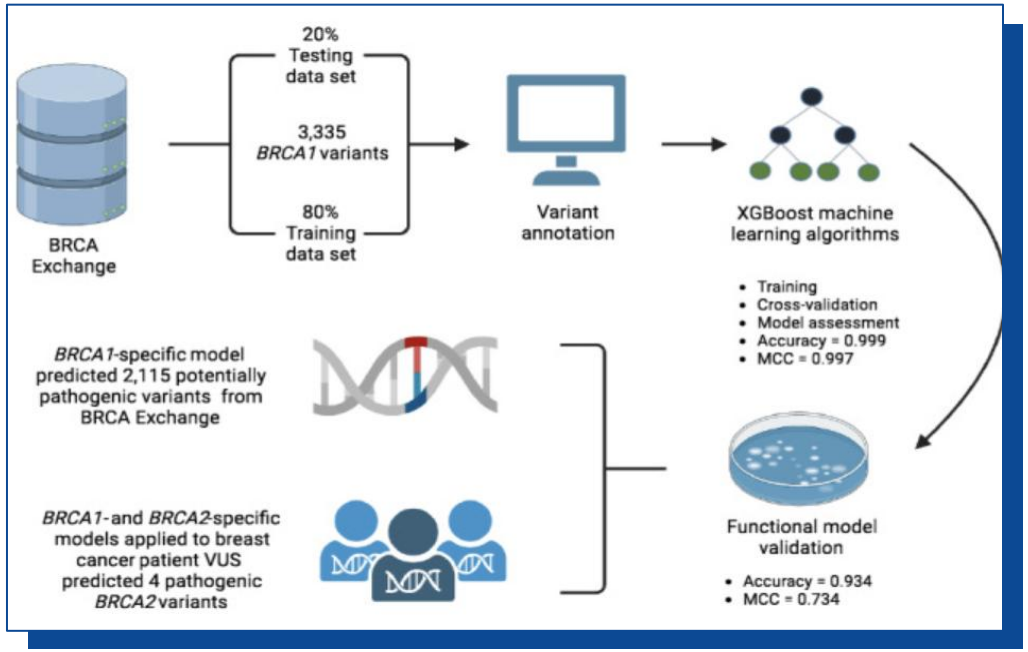

图1.3 XGBoost 模型预测BRCA1概率工作流程[5]

机器学习面临的挑战

- ✓ 基因组数据具有典型的“高维度、低样本量”（HDLSS）特征
- ✓ 数据中存在大量噪声和冗余信息
- ✓ 部分亚型特征相似，界限模糊，且存在样本不平衡问题

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[6]。

Lee 等人则进一步拓展了图神经网络的应用，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 AttentionNetwork），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[7]。

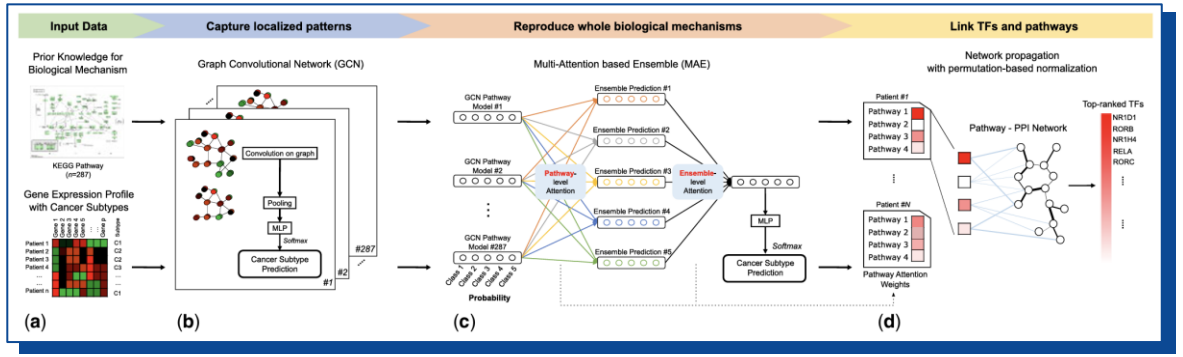

图1.4 多重注意力图卷积网络模型的工作流程[7]

机器学习可挖掘BRCA相关基因表达潜在模式，高效识别乳腺癌分子亚型与致病风险，助力乳腺癌早筛与个体化诊疗

[5] Khandakji M, Habish H H A, Abdulla N B S, et al. BRCA1-specific machine learning model predicts variant pathogenicity with high accuracy[J]. Physiological Genomics, 2023, 55(8): 315-323.
[6] 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.
[7] 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. Bioinformatics36(12), 3818–3824.

1.3 研究目标

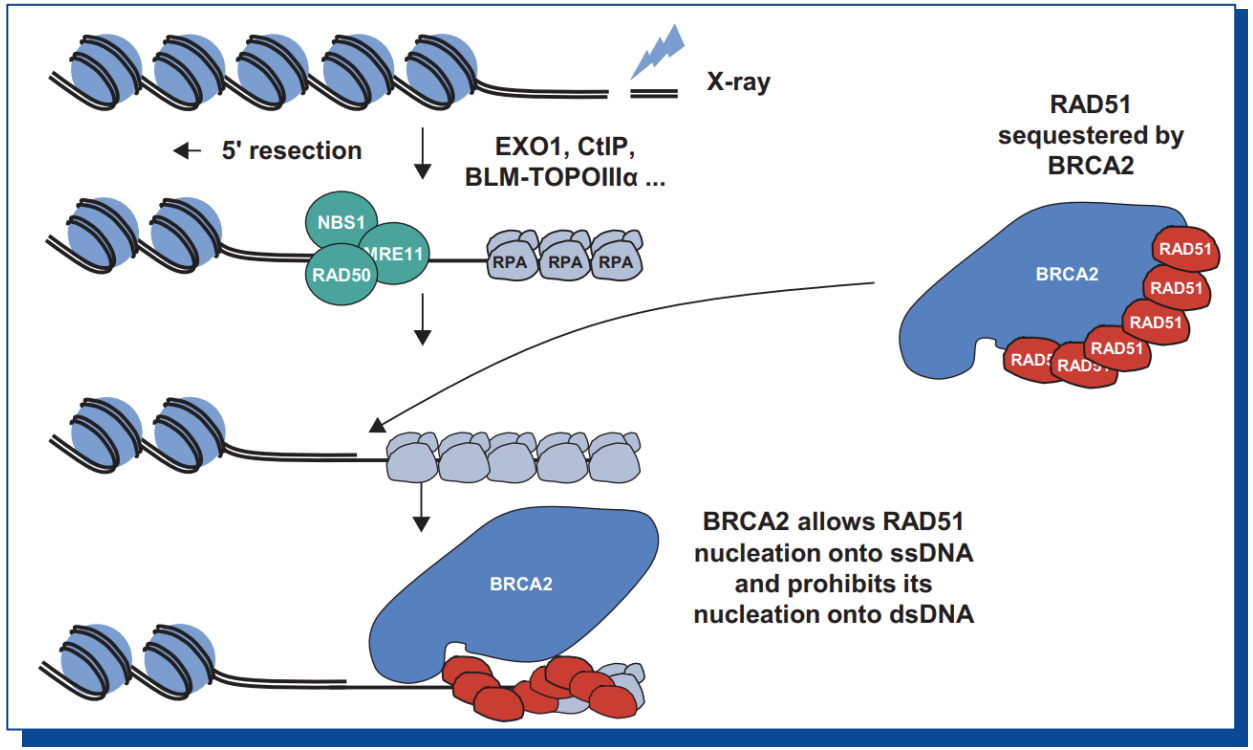

图1.5 BRCA2在通过同源重组修复DNA双链断裂中的作用[8]

核心研究目标

- ✓ 系统评估：全面评估和比较五种主流机器学习模型（LightGBM, XGBoost, MLP, CNN, SVM）在PAM50亚型分类任务上的性能。
- ✓ 探究方法：深入分析数据预处理、特征选择和超参数优化对模型最终性能的关键影响。
- ✓ 总结优劣：系统总结各模型在处理基因拷贝数变异（CNV）数据时的优势、劣势及适用场景。

核心焦点

- ✓ **各机器学习模型的性能指标对比分析**
- ✓ **超参数优化策略及其对模型效能的影响**
- ✓ **面向基因组学数据的特有方法学考量**
- ✓ **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**

[8] Stoppa-Lyonnet D. The biological effects and clinical implications of BRCA mutations: where do we go from here?[J]. European Journal of Human Genetics, 2016, 24(1): S3-S9.

02

PART TWO

01 研究背景与问题提出

02 **数据获取与方法应用**

03 模型实现与综合评估

04 实验总结与结论思考

2.1 数据来源

图2.1 TCGA-BRCA乳腺癌基因组图谱收集项目[9]

Gene Symbol	TCGA-3C-AAAU-01	TCGA-3C-AALI-01	TCGA-3C-AALJ-01	TCGA-3C-AALK-01	TCGA-4H-AAAK-01	TCGA-5L-AAT0-01	TCGA-5L-AAT1-01	TCGA-5T-A9QA-01	TCGA-A1-A0S
ACAP3	0	-1	-1	0	0	0	0	-1	0
ACTRT2	0	-1	-1	0	0	0	0	-1	0
AGRN	0	-1	-1	0	0	0	0	-1	0
ANKRD65	0	-1	-1	0	0	0	0	-1	0
ATAD3A	0	-1	-1	0	0	0	0	-1	0
ATAD3B	0	-1	-1	0	0	0	0	-1	0
ATAD3C	0	-1	-1	0	0	0	0	-1	0
AURKAIP1	0	-1	-1	0	0	0	0	-1	0
B3GALT6	0	-1	-1	0	0	0	0	-1	0
C1orf159	0	-1	-1	0	0	0	0	-1	0
C1orf170	0	-1	-1	0	0	0	0	-1	0
C1orf222	0	-1	-1	0	0	0	0	-1	0
C1orf233	0	-1	-1	0	0	0	0	-1	0
C1orf86	0	-1	-1	0	0	0	0	-1	0
CALML6	0	-1	-1	0	0	0	0	-1	0
CCNL2	0	-1	-1	0	0	0	0	-1	0
CDK11A	0	-1	-1	0	0	0	0	-1	0
CDK11B	0	-1	-1	0	0	0	0	-1	0
CPSF3L	0	-1	-1	0	0	0	0	-1	0
DDX11L1	0	-1	-1	0	0	0	0	-1	0
DVL1	0	-1	-1	0	0	0	0	-1	0
FAM132A	0	-1	-1	0	0	0	0	-1	0

图2.2基因拷贝数变异 (CNV) 数据（部分）

使用数据：TCGA-BRCA 项目

TCGA-BRCA项目[1]（The Cancer Genome Atlas - Breast Invasive Carcinoma）是美国癌症基因组图谱（TCGA）计划下的一个核心子项目，旨在系统解析乳腺浸润性癌（BRCA）的分子机制。**该项目通过多组学平台（包括基因组测序、转录组分析、DNA甲基化、拷贝数变异、miRNA表达等）**对超过**1000例乳腺癌样本**及其对应的正常组织进行了深度分析，涵盖了**分子分型、驱动基因突变、信号通路重编程及临床特征关联**等多维度信息。

模型特征：基因拷贝数变异 (CNV) 数据

数据格式：本研究采用的数据是经过 **GISTIC2.0 算法** 处理并**阈值化**的基因水平拷贝数估计。

- ✓ **-2: 纯合性缺失**
- ✓ **1: 低水平扩增/增益**
- ✓ **-1: 杂合性缺失**
- ✓ **2: 高水平扩增**
- ✓ **0: 正常二倍体 (Diploid)**

[1] The Cancer Genome Atlas Network. 2012. Comprehensive molecular portraits of human breast tumours. Nature 490(7418), 61–70.
[9] Tomczak, K., Czerwińska, P., & Wiznerowicz, M. 2015. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. Contemporary Oncology 19(1A), A68– A77.

2.2 数据预处理

```
# 筛选肿瘤样本 (sample code 01-09)
phenotype_data_raw$sample_code_char <- substr(phenotype_data_raw$sampleID, 14, 15)
phenotype_tumor_samples <- phenotype_data_raw %>%
  filter(as.integer(sample_code_char) >= 1 & as.integer(sample_code_char) <= 9)

# 提取并清洗亚型标签
phenotype_selected_subtypes <- phenotype_tumor_samples %>%
  select(sampleID, label = !!sym(subtype_column_name)) %>%
  filter(!is.na(label) & label != "" & label != "null")

# 移除样本数过少的稀有亚型 (例如，少于7个)
subtype_counts <- table(phenotype_selected_subtypes$label)
rare_subtypes <- names(subtype_counts[subtype_counts < 7])
phenotype_selected_subtypes <- phenotype_selected_subtypes %>%
  filter(!(label %in% rare_subtypes))

# 移除低方差特征
final_data_for_ml <- merged_data_subtypes %>% select(-sampleID)
numeric_feature_columns_data <- final_data_for_ml[, setdiff(names(final_data_for_ml), "label")]
feature_vars <- apply(numeric_feature_columns_data, 2, var, na.rm = TRUE)
constant_features <- names(feature_vars[feature_vars == 0 | is.na(feature_vars)])
final_data_for_ml <- final_data_for_ml %>% select(-all_of(constant_features))

# 基于高方差进行激进的特征预过滤
TARGET_NUM_FEATURES_PREFILTER <- 150 # 目标保留的特征数量
feature_data_only <- final_data_for_ml[, setdiff(names(final_data_for_ml), "label")]
feature_variances <- apply(feature_data_only, 2, var, na.rm = TRUE)
top_variance_features <- names(sort(feature_variances, decreasing =
TRUE)[1:TARGET_NUM_FEATURES_PREFILTER])
```

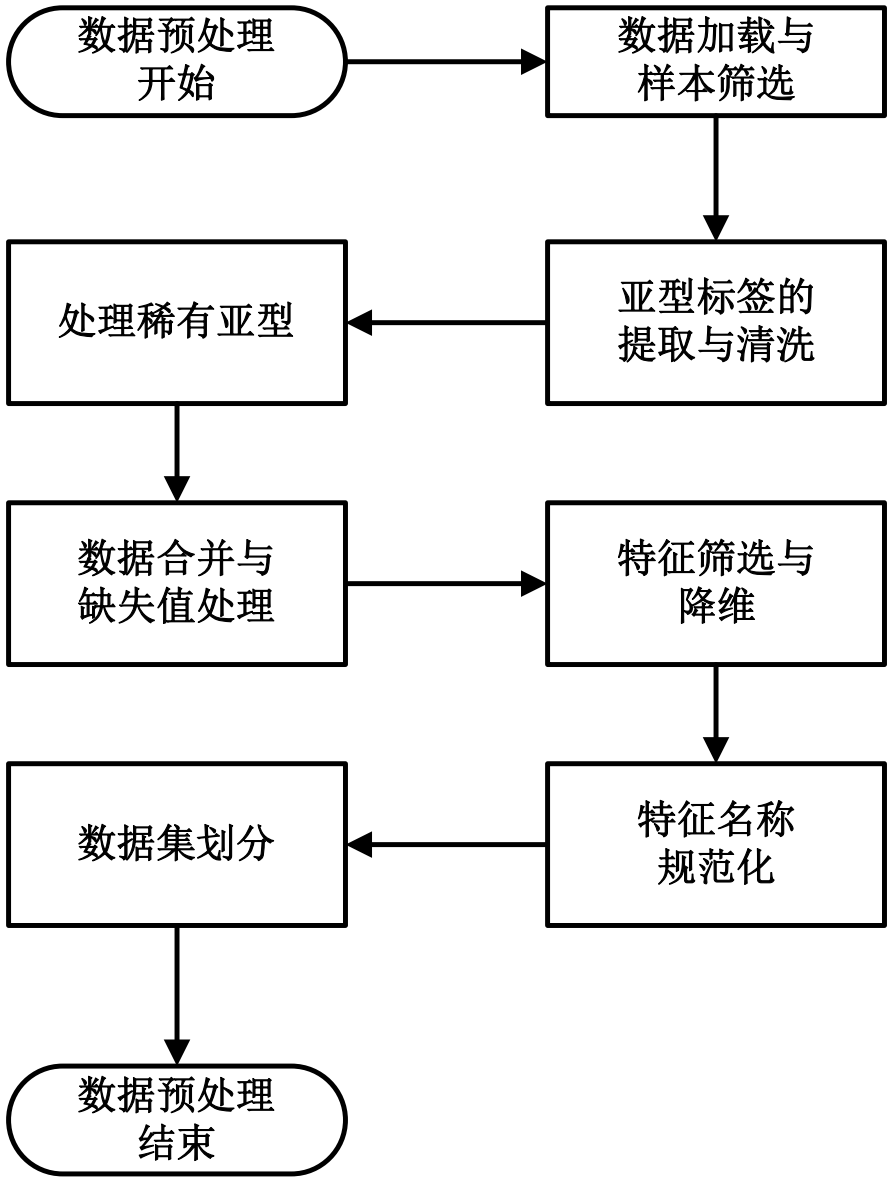

图2.3 数据预处理流程

2.2 数据预处理

```
# 规范化特征列名
features_to_clean <- setdiff(colnames(final_data_for_ml), "label")
cleaned_feature_names <- make.names(features_to_clean, unique = TRUE)
colnames(final_data_for_ml)[-1] <- cleaned_feature_names

# 设置随机种子
set.seed(42)

# 使用分层抽样进行数据划分 (60%训练, 20%验证, 20%测试)
# 划分出20%的测试集
test_indices <- createDataPartition(final_data$label, p = 0.20, list = FALSE)
test_set <- final_data[test_indices, ]
train_validation_set <- final_data[-test_indices, ]

# 从剩余的80%数据中划分出验证集
validation_indices_in_tv <- createDataPartition(train_validation_set$label, p = 0.25, list =
FALSE)
validation_set <- train_validation_set[validation_indices_in_tv, ]
train_set <- train_validation_set[-validation_indices_in_tv, ]

output_dir_splits <- "data_splits"
if (!dir.exists(output_dir_splits)) {
  dir.create(output_dir_splits)
}
saveRDS(train_set, file = file.path(output_dir_splits, "train_set_subtypes.rds"))
saveRDS(validation_set, file = file.path(output_dir_splits, "validation_set_subtypes.rds"))
saveRDS(test_set, file = file.path(output_dir_splits, "test_set_subtypes.rds"))
```

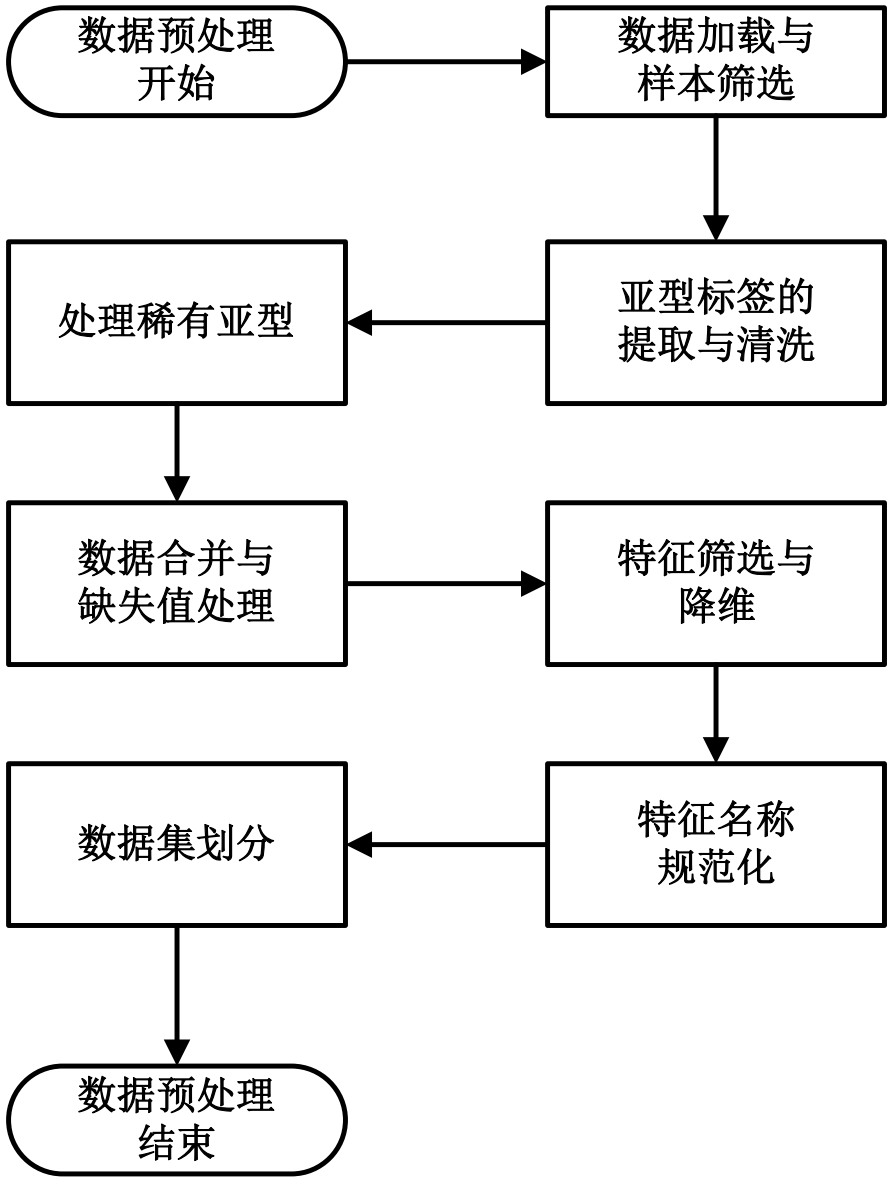

图2.3 数据预处理流程

2.3 机器学习模型概述——LightGBM[10]

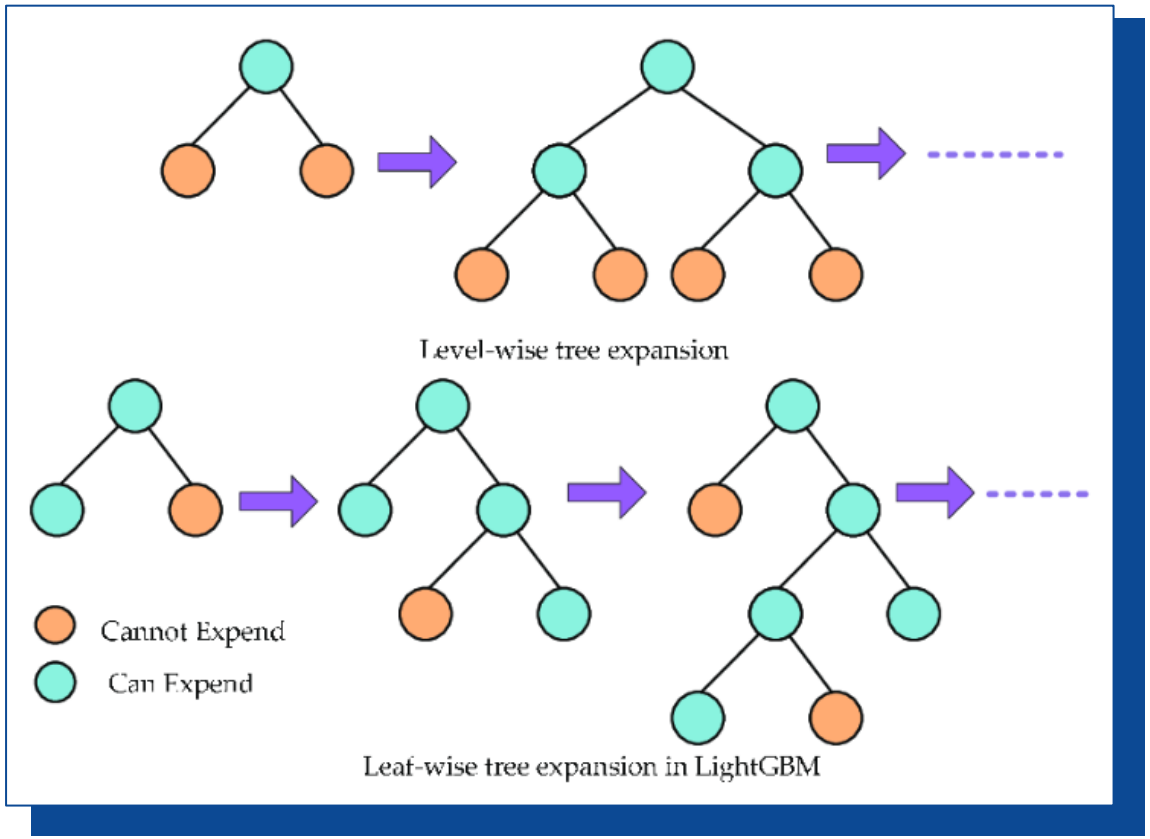

图2.4 LightGBM

算法原理

基于梯度提升决策树（GBDT），但采用“叶子优先”增长策略，并结合梯度单侧采样（Gradient-Based One-Side Sampling, GOSS）与特征绑定（Exclusive Feature Bundling, EFB），以提高训练效率和内存利用率。

算法应用

用于乳腺癌诊断中，在Wisconsin数据集上取得优于GBM和XGBoost的表现。LightGBM 模型的准确率、召回率、AUC 和精确率为95.3%、94.8%、0.987、95.5%。[11]

[10] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Advances in Neural Information Processing Systems 30.
[11] Akbulut S, Cicek IB, Colak C. Classification of Breast Cancer on the Strength of Potential Risk Factors with Boosting Models: A Public Health Informatics Application. Med Bull Haseki. 2022 Jun;60(3):196-203.

2.3 机器学习模型概述——XGBoost [12]

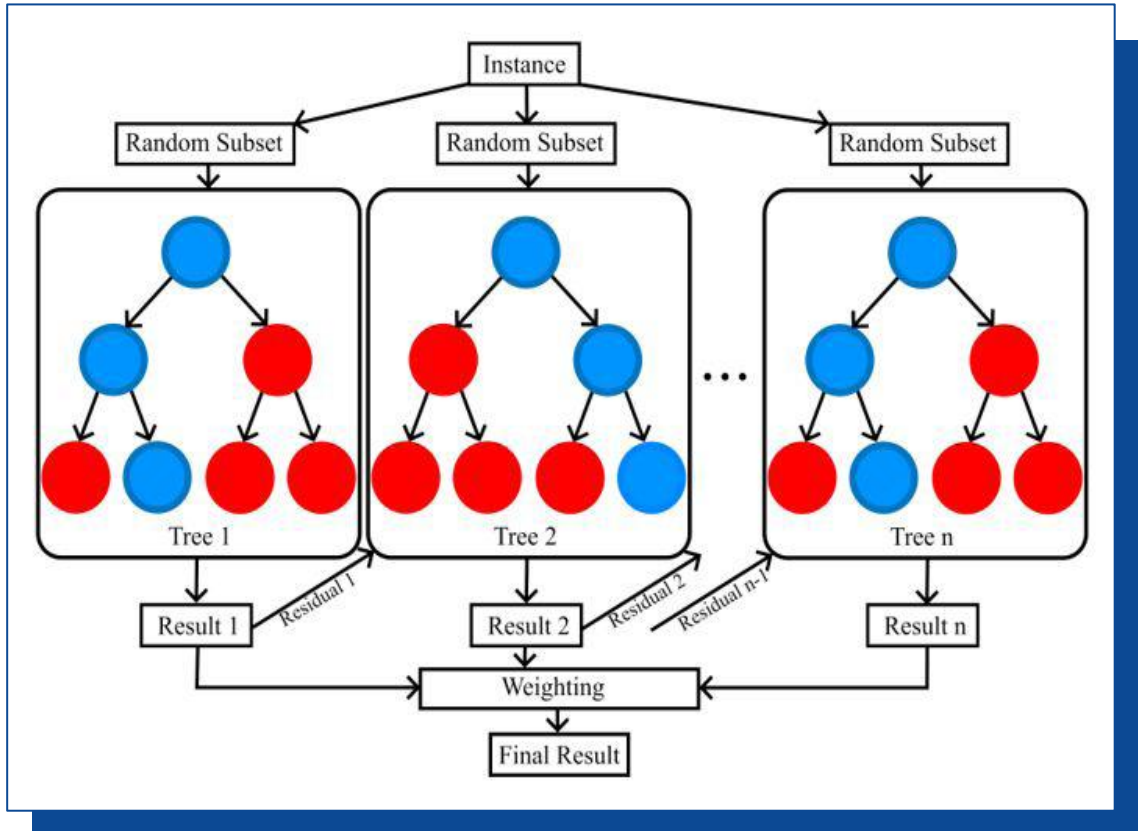

图2.5 XGBoost

算法原理

增强型梯度提升树，使用二阶导数信息、正则化项，支持自动处理缺失值、类别稀疏，树构建采用预排序/直方图方法。正则化的学习目标如下：

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

算法应用

Khandakji & Mifsud开发 BRCA2 基因特异 XGBoost 模型预测 VUS（变异不确定性），模型准确率高达 99.9%，并通过自带因子功能识别重要致病特征。研究基于二阶梯度信息的增强型 GBDT，支持缺失值处理、正则化目标和特征重要性评估；用于构建高性能 pathogenicity 分类器。 [13]

[12] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM.
[13] Khandakji, M.N., Mifsud, B. A gene-specific XGBoost model to predict pathogenicity of BRCA2 variants. PeerJ. 2022;10:e13631.

2.3 机器学习模型概述——SVM[14-15]

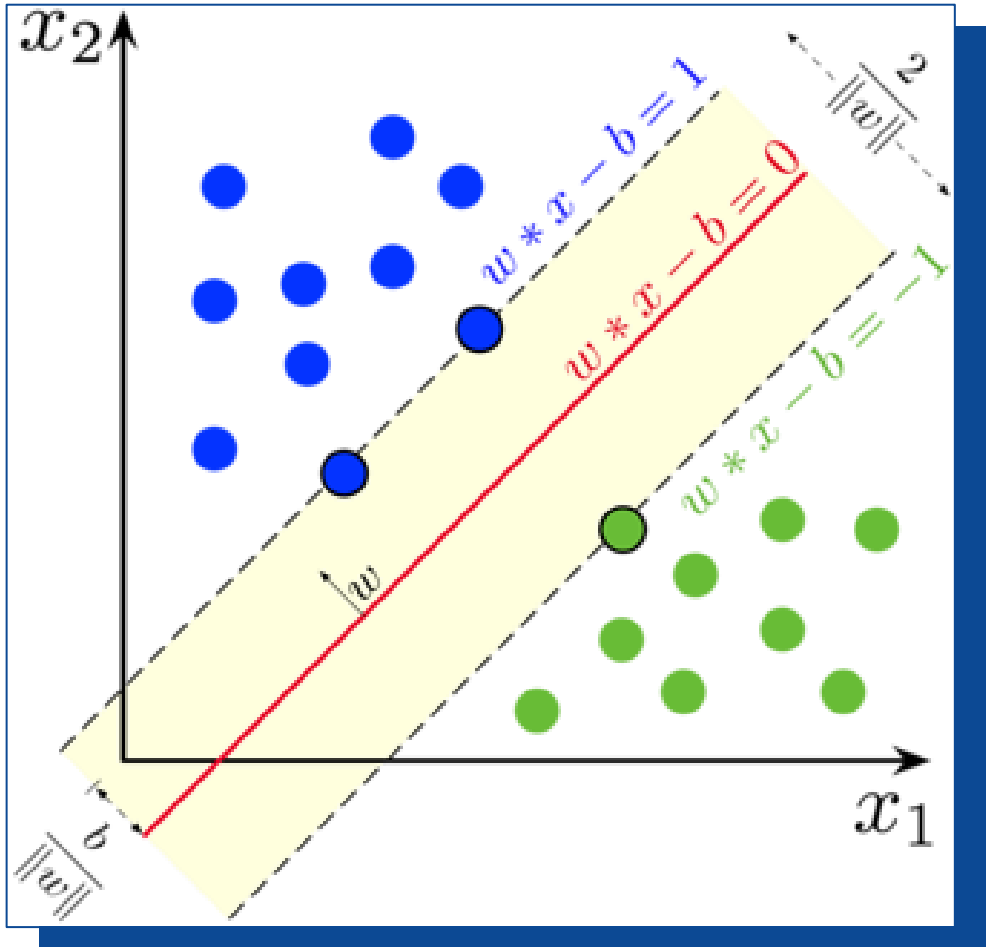

图2.6 SVM

算法原理

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。

算法应用

Zhao 等使用 SVM 构建 BRCA1-like 复制分型分类器，利用基因组拷贝数谱区分 BRCA1 型乳腺癌，训练集 AUC=1.00，验证集 AUC≈0.75。研究采用核方法映射至高维空间，通过最大化类别间隔实现分类；适合 HDLSS 数据，通过 ROC 曲线评估模型泛化性。[16]

[14] Vapnik, V. N., & Chervonenkis, A. Y. (1964). Об одном классе алгоритмов обучения распознаванию образов. Автоматика и телемеханика, 25(6), 937.
[15] Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. Machine Learning, 20(3), 273-297
[16] Zhao, X., Guo, T., Lu, Y., Liu, J. Development of a support vector machine BRCA1-like classifier for breast cancer based on copy number data. Physiol Genomics. 2023;55(4):112–119.

2.3 机器学习模型概述——MLP[17]

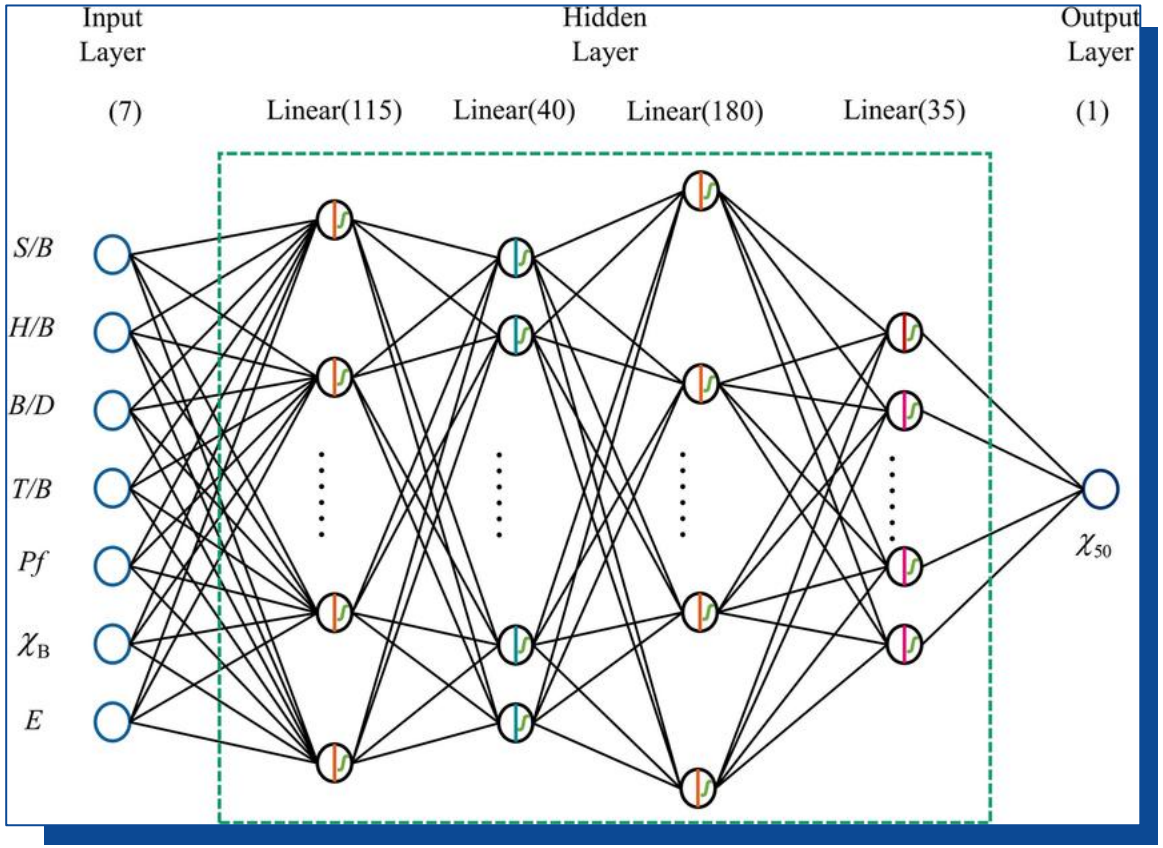

图2.7 MLP

算法原理

多层感知器（MLP）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。

算法应用

Rajpal 等提出基于自动编码器+MLP 的两阶段深度学习模型，对 TCGA-BRCA 基因表达数据进行子型分类（Basal, HER2, LumA, LumB），10 折交叉验证平均准确率达到 90.7%。研究使用 AE 降维将数万维基因表达映射至低维特征，再通过 fully-connected MLP 完成分类，适应高维非线性映射任务。[18]

[17] McCulloch, W. S., & Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics, 5, 115–133.
[18] Rajpal, S., Kumar, V., Agarwal, M., Kumar, N. Deep learning-based model for breast cancer subtype classification. arXiv. 2021. arXiv:2111.03923.

2.3 机器学习模型概述——CNN[20]

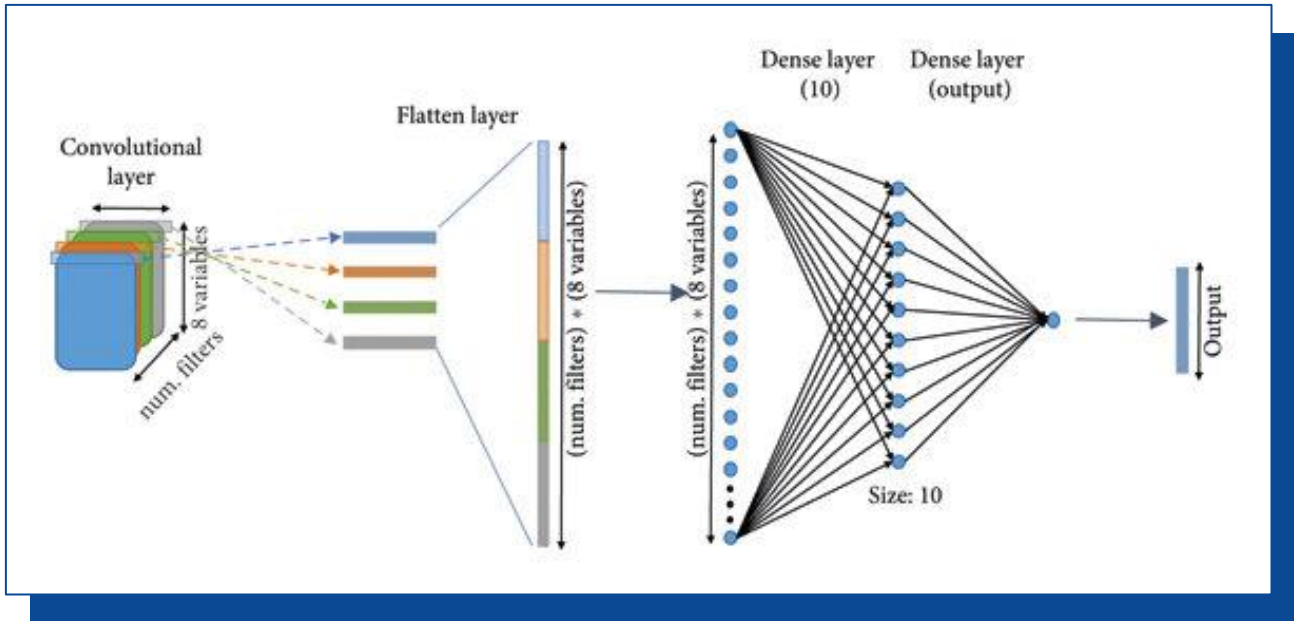

图2.8 Architecture of one-dimensional convolutional neural network (1DCNN) [19]

算法原理

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。

算法应用

Elbashir 等构建 bio-inspired CNN 结合优化算法 EOSA，处理 TCGA-BRCA 基因表达谱作为“图像”进行癌症组织识别，模型准确率达 98.3%。研究将基因表达向量重塑为二维图像输入 CNN，利用卷积层提取局部模式，EOSA 优化权重，以识别肿瘤与正常样本。[21]

[19] Aceves-Fernández M A, Domínguez-Guevara R, Pedraza-Ortega J C, et al. Evaluation of key parameters using deep convolutional neural networks for airborne pollution (PM10) prediction[J]. Discrete Dynamics in Nature and Society, 2020, 2020(1): 2792481.
[20] Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, 36(4), 193–202.
[21] Elbashir, A.A., et al. A bio-inspired convolution neural network architecture for automatic breast cancer detection using gene expression data. Sci Rep. 2023;13:41731. doi:10.1038/s41598-023-41731-z

03

PART THREE

01 研究背景与问题提出

02 数据获取与方法应用

03 模型实现与综合评估

04 实验总结与结论思考

3.1 LightGBM [10] 实现与评估

表3 LightGBM 模型在测试集上的性能指标

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

表4 LightGBM 最优超参数配置

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

表5 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。

```
# 通过交叉验证寻找最佳参数和轮数
lgb_cv_params_list <- list(
  objective = "multiclass",
  metric = "multi_logloss",
  num_class = num_classes,
  # ... (learning_rate, num_leaves等参数来自current_params_row) ...
)
lgb_cv_run <- lgb.cv(
  params = lgb_cv_params_list,
  data = dtrain_lgb,
  nrounds = 200,
  nfold = 5,
  early_stopping_rounds = 30,
  stratified = TRUE,
  verbose = -1
)
best_nrounds_lgb <- lgb_cv_run$best_iter

# 使用最佳参数训练最终模型
final_lgb_params_list <-
final_lgb_model <- lgb.train(
  params = final_lgb_params_list,
  data = dtrain_lgb,
  nrounds = best_nrounds_from_cv_lgb,
  valids = list(eval = dvalidation_lgb),
  record = TRUE
)
```


[10] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Advances in Neural Information Processing Systems 30.

3.2 XGBoost [12]实现与评估

表6 XGBoost 模型在测试集上的性能指标

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

表7 XGBoost 最优超参数配置

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

表8 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。

```
# 通过交叉验证寻找最佳参数和轮数 (在参数搜索循环内)
xgb_cv_params_list <- list(
  objective = "multi:softprob", # 多分类概率输出
  eval_metric = "mlogloss",    # 评估指标
  num_class = num_classes,
  # ... (eta, max_depth, subsample等参数来自current_params) ...
)
xgb_cv_run <- xgb.cv(
  params = xgb_cv_params_list,
  data = dtrain,
  nrounds = 200,
  nfold = 5,
  early_stopping_rounds = 20,
  stratified = TRUE,
)
best_nrounds <- xgb_cv_run$best_iteration

# 使用最佳参数训练最终模型
final_params_list <-
final_xgb_model <- xgb.train(
  params = final_params_list,
  data = dtrain,
  nrounds = best_nrounds_from_cv,
  watchlist = list(train = dtrain, eval = dvalidation),  print_every_n = 10
)
```


[12] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM.

3.3 支持向量机 (SVM) [14-15]实现与评估

表9 SVM模型在测试集上的性能指标

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

表10 SVM最优超参数配置

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

表11 SVM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。

```r
# 准备训练控制参数和超参数网格
train_control_svm <- trainControl(
  method = "cv",
  number = 5,
  summaryFunction = multiClassSummary,
  classProbs = TRUE,
  verboseIter = TRUE
)

# 定义要搜索的超参数范围
tune_grid_svm_radial <- expand.grid(
  sigma = 10^seq(-5, -2, length.out = 4),
  C = 2^seq(-10, -6, length.out = 5)
)
```

```r
# 训练模型
set.seed(123)
caret_svm_model <- train(
  label ~ .,
  data = train_set,
  method = "svmRadial",
  trControl = train_control_svm,
  tuneGrid = tune_grid_svm_radial,
  metric = "Mean_F1",
  preProcess = c("center", "scale")
)
```


[14] Vapnik, V. N., & Chervonenkis, A. Y. (1964). Об одном классе алгоритмов обучения распознаванию образов. Автоматика и телемеханика, 25(6), 937.
[15] Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. Machine Learning, 20(3), 273-297

3.4 多层感知器 (MLP) [17]实现与评估

表12 MLP模型在测试集上的性能指标

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

表13 MLP最优超参数配置

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

表14 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）。

```
# 标准化特征
train_mean <- apply(x_train, 2, mean)
train_sd <- apply(x_train, 2, sd)
x_train_scaled <- scale(x_train, center = train_mean, scale = train_sd)
x_val_scaled <- scale(x_val, center = train_mean, scale = train_sd)
x_test_scaled <- scale(x_test, center = train_mean, scale = train_sd)

# One-Hot编码标签
y_train_numeric <- as.numeric(train_set$label) - 1
y_train_one_hot <- to_categorical(y_train_numeric, num_classes = num_classes)
# ... (对 y_val, y_test 做同样处理) ...
```

```
# 在循环中定义和编译模型
# current_params 来自超参数网格 (param_grid_mlp)
model_mlp <- keras_model_sequential() %>%
  layer_dense(units = current_params$units1, activation = "relu", input_shape = ncol(x_train_scaled)) %>%
  layer_batch_normalization() %>%
  layer_dropout(rate = current_params$dropout_rate1) %>%
  layer_dense(units = current_params$units2, activation = "relu") %>%
  layer_batch_normalization() %>%
  layer_dropout(rate = current_params$dropout_rate2) %>%
  layer_dense(units = num_classes, activation = "softmax") # 输出层
model_mlp %>% compile(
  optimizer = optimizer_adam(learning_rate = current_params$learning_rate),
  loss = "categorical_crossentropy",
  metrics = c("accuracy")
)
```


[17] McCulloch, W. S., & Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics, 5, 115–133.

3.5 卷积神经网络 (CNN) [20]实现与评估

表15 CNN模型在测试集上的性能指标

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

表16 CNN最优超参数配置

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

表17 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）。

```
# 将2D数据 (样本 x 特征) 塑形为3D张量 (样本 x 特征 x 通道) 以适应1D CNN
num_features <- ncol(x_train_scaled)
x_train_reshaped <- array_reshape(x_train_scaled, c(nrow(x_train_scaled), num_features, 1))
x_val_reshaped <- array_reshape(x_val_scaled, c(nrow(x_val_scaled), num_features, 1))
x_test_reshaped <- array_reshape(x_test_scaled, c(nrow(x_test_scaled), num_features, 1))

# 定义和编译模型 (此处展示一个示例结构)
# current_params 来自超参数网格 (param_grid_cnn)
input_shape_cnn <- c(num_features, 1)
model_cnn <- keras_model_sequential() %>%
  layer_conv_1d(
    filters = current_params$filters1,
    kernel_size = current_params$kernel_size1,
    activation = "relu",
    input_shape = input_shape_cnn
  ) %>%
  layer_batch_normalization() %>%
  layer_max_pooling_1d(pool_size = current_params$pool_size1) %>%
  layer_dropout(rate = current_params$dropout_cnn) %>%
  layer_global_average_pooling_1d() %>% # 全局池化层替代Flatten
  layer_dense(units = current_params$dense_units, activation = "relu") %>%
  layer_dropout(rate = current_params$dropout_dense) %>%
  layer_dense(units = num_classes, activation = "softmax") # 输出层
model_cnn %>% compile(...)
```


[20] Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, 36(4), 193–202.

3.6 多模型综合评估

表18 PAM50亚型分类的代表性模型测试集性能比较摘要

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
LightGBM	0.6176	0.6080	0.6380	0.6155	lr=0.05, nl=31, md=4, ff=0.8, bf=0.7, nrounds=102
XGBoost	0.6569	0.7230	0.6460	0.6614	eta=0.075, md=3, sub=0.6, col=0.7, nrounds=78
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
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
SVM	0.4510	0.4767	0.4456	0.4467	sigma=0.001, C=0.0116, RBF 核, 500 特征

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。

3.6 多模型综合评估

图3.4 各模型在PAM50测试集上的指标表现

3.7 软件平台

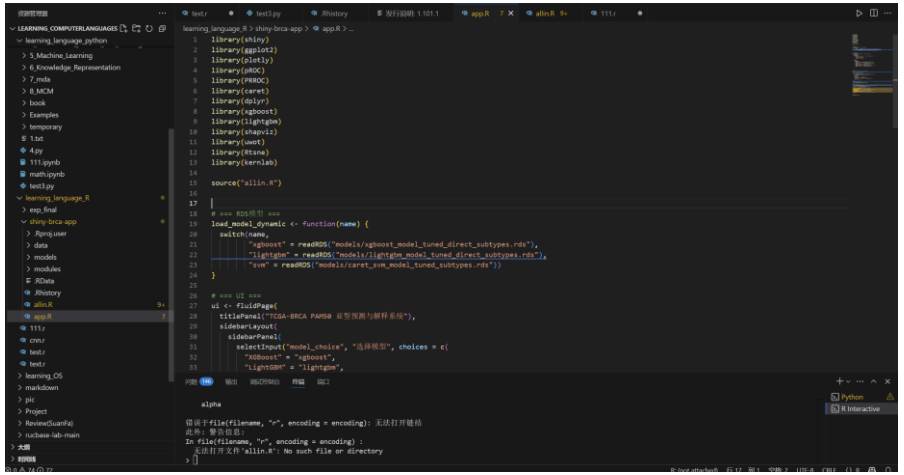

图3.2 软件平台代码

图3.3清晰地揭示了模型在预测不同乳腺癌 PAM50 亚型时，对各个基因特征的依赖程度，强调了像 GRB7、IKZF3、ERBB2 等基因在区分 Basal-like、HER2-enriched 和 Luminal A 亚型方面具有显著贡献。

图3.3 软件平台展示（部分）

04

PART FOUR

01 研究背景与问题提出

02 数据获取与方法应用

03 模型实现与综合评估

04 **实验总结与结论思考**

4.1 性能差异深度分析

梯度提升模型为何胜出?

- ✓ **模型与数据匹配度高**：树模型天然适合处理具有复杂交互作用和非线性关系的表格型数据。
- ✓ **对数据预处理要求低**：对特征的尺度不敏感，无需严格的标准化。
- ✓ **强大的内置正则化**：XGBoost和LightGBM通过控制树的复杂度、采样等多种方式有效防止过拟合。

神经网络模型为何挣扎?

- ✓ **数据量依赖性**：深度学习模型通常需要大规模的训练数据才能充分学习其众多参数，避免过拟合。本研究数百级别的样本量可能不足。
- ✓ **CNN对输入结构的敏感性**：1D-CNN的性能依赖于输入序列中特征的排列顺序。当前基于方差选择的基因排列，可能无法为CNN提供易于学习的局部模式。

SVM为何表现平平?

- ✓ **对特征质量的敏感性**：SVM的性能高度依赖于输入特征的质量。即使增加到500个特征，如果其中包含大量噪声或冗余信息，也可能损害而不是提升性能。
- ✓ **对核函数与参数的敏感性**：RBF核的参数sigma和C的选择至关重要，当前的网格搜索可能未能找到全局最优解。

4.2 未来研究方向与策略建议

数据层面

- ✓ **多组学数据整合：**融合基因表达谱、DNA甲基化、等数据，构建更全面的肿瘤分子画像，有望突破单一CNV数据的局限性。
- ✓ **数据增强技术：**对于样本量不足的问题，可探索使用生成对抗网络等技术来生成高质量的合成数据扩充训练集。
- ✓ **多模态技术**：可以加入生化指标、图像样本进行训练，提升训练效果。

方法层面

- ✓ **高级特征工程与选择：**应用更复杂的特征选择算法（如LASSO），或结合生物学先验知识（如通路信息）来筛选特征。
- ✓ **神经网络深度优化：**为CNN设计更有效的输入表示（如将基因按染色体位置排序）；为MLP探索更深或更先进的网络架构（如残差网络）。

解释性与应用层面

- ✓ **强化模型可解释性：**对表现最佳的XGBoost模型，应用SHAP等XAI工具，深入理解驱动亚型分类决策的关键基因及其贡献模式。
- ✓ **处理类别不平衡与亚型模糊性：**针对Luminal A/B等混淆严重的亚型，采用代价敏感学习或更高级的采样技术（如SMOTE变体）进行针对性优化。

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 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
	文浩名	中国地质大学
	刘星雨	北京体育大学

汇报人：张鲴沣 | 汇报时间：2025/06/22