# 摘 要

胃癌是全球常见的消化系统恶性肿瘤，其发病率和死亡率逐年上升．早期胃癌治愈率较高，但易被误诊，耽误治疗时机，造成更严重后果．因此，科学有效的预测方法对于胃癌的诊断至关重要.机器学习技术的发展为检测胃癌风险提供了简单有效的方式，有助于降低胃癌的发病率．

本文从TCGA数据库中获取了胃癌患者的基因表达数据，并对原始数据集进行了预处理．结合了多个特征筛选方法来对特征基因进行筛选，并构建集成分类学习器来对最优特征子集的分类性能进行评价.利用患癌样本的临床信息进行生存分析，结果表明所筛选出的特征基因的表达量与胃癌患者的生存率密切相关．本文主要工作有三个方面．

(1)胃癌特征基因筛选．本研究采用基因表达差异分析方法筛选出1300个差异基因，随后采用最小冗余最大相关算法选取最小冗余且与类别最相关的前300个特征基因．在此基础上，采用基于参数寻优的SVM-RFECV-PSO算法进行特征筛选，最终得到了19个特征基因用于建模分析．

(2)在筛选过的数据上构建胃癌预测模型．首先，使用随机森林和XGBoost算法分析得到正确率和AUC值分别为95.12%、94.31%和0.9923、0.9867．然后，使用结合多种学习算法的Stacking模型进行预测，准确率为98.52%、AUC值为0.9979，筛选的特征基因具有很好的区分度，Stacking模型预测效果最佳．

(3)胃癌患者的生存分析．本文筛选出的19个特征基因，在检索后发现至少6个与人体消化系统或胃癌预后状况有关．用基因CLDN7、GCNT4、BAIAP2L2和ALDH3A1的表达数据与临床信息结合展开生存分析，结果表明这些基因对胃癌患者的生存状况有显著影响．

本文研究了胃癌基因表达数据的特征选择和集成学习模型的构建，为胃癌早期诊断提供了新思路和方法．同时，基于特征基因的临床数据分析也为胃癌患者的治疗和生存提供了参考．该研究对胃癌早期诊断和及时治疗具有一定的价值．

**关键词：**胃癌，特征基因筛选，Stacking，生存分析

# 目 录

# 第1章 引言

## 1.1 研究背景及意义

癌症是一类恶性肿瘤的统称，其细胞可以在人体内无限制地增殖和扩散，最终可能会扩散到身体其他部位[1]. 当前，癌症的患病率与致死率持续不断上升，这对人类的身体和生命构成了巨大威胁. 胃癌是全球范围内常见的消化系统恶性肿瘤，2020年全球癌症报告显示胃癌已成为世界第五大最常被诊断的癌症和第四大癌症死亡原因，在每年新增100多万病例中，约有76.9万人死亡，在东亚地区分布较为集中，我国也是胃癌的高发地区[2]. 近年来我国胃癌发病及病死率逐年递增，报告显示我国约占全球胃癌总发病率的49%，占全球胃癌年新发病人数的44%[3]. 许多胃癌患者确诊时已是晚期，没有手术机会，这部分患者的疗效及生存期差. 随着分子生物学的研究，医疗检测技术的进步，部分患者经过早诊早治，生存期得到延长，但统计数据显示我国晚期胃癌患者5年生存率仍低于30%[4]，现状仍不容乐观. 机器学习能够更简便更有效地检测出人们是否有患癌的危险，从而减少癌症的发生率，提升人们的生活品质[5].

由于基因检测技术不断进步，采用基因检测技术对胃癌的早期诊断越来越受到重视. 在每一个样本上，都记录了全部可测基因的表达量，但是只有少量的基因与样本类别有关联，它们含有大量关于样本分类的信息[6]. 所以，在癌症基因表达分析中，选择具有判别能力的特征信息基因是至关重要的. 这不仅涉及到建立有效的分类模型，同时是寻找可能的基因标志物和药物作用靶点的关键工具[7]. 由于基因的表达数据具有“小样本、高维度”的特点，从其中挖掘出有鉴别能力的特征基因，是揭示基因和肿瘤的关系、深入了解疾病的发生机制和提高疾病的临床诊断精度的关键步骤[8].

特征基因的选择就是从原有的基因集合中选出最优特征子集. 特征筛选方法能够剔除不需要的特征，降低数据集的维度，是构建高效、通用且具有较强预测性能的分类器的关键. 然而，传统的特征选择方法大多建立在统计学的基础上，而忽视了生物学中的一些知识. 实际上，只有一小部分基因确实与样本的类型有

关. 因此，寻找与样本表型相关的基因，即特征基因，是癌症基因表达数据分析的核心问题，也是胃癌特征选择的难点. 因此，对已有的特征选择算法进行改进，使其具有更好的分类性能，已经成为胃癌分类研究中的一个重要课题.

通常在进行特征选择之后，需要利用分类算法对选中的最优特征子集展开分类，并对其进行评估，所采用分类器的性能会对胃癌分类的精度产生直接影响，因此，研究具有良好分类能力的分类模型成为重点. 然而，单一的分类器模型存在过拟合、泛化性不强和预测准确率不稳定等问题. 因此，可以选择将多个单分类器集成为一个强分类器，这个强分类器被称为集成分类器，它能够有效地改善分类模型的性能，同时可以更好地评估特征选取的性能.

综上所述，特征基因选择算法在胃癌分类研究中扮演者重要角色. 目前，研究人员正在致力于寻找更准确和高效的特征选择算法，改进现有的算法也是一个关键的研究方向. 此外，利用集成分类器不但能够提高胃癌分类的精度，而且可以对特征选择算法进行更综合的评价，因而利用集成分类器进行优化是胃癌分类研究中的一个重要问题.

## 1.2 国内外研究现状

选择特征基因是在生物信息学中分析基因表达数据时不可或缺的步骤之一，尤其在癌症研究中备受关注. 随着高通量技术的广泛应用，基因表达数据的数量和质量不断提高，为特征基因选择提供了更多的数据资源和分析工具，同时也推动了特征基因选择方法的不断创新和优化. 在癌症特征基因选择和分类研究方面，统计学习、机器学习等领域近年来取得显著进展，成为生物信息学领域的核心技术.

对于特征基因的筛选，当前常用的统计学方法有：T检验、ANOVA分析、皮尔逊相关系数、互信息、岭回归、LASSO回归以及Chu G等人[9]在2001年提出的的SAM方法. 除此之外，在机器学习中还涌现出了很多可以对高维数据进行处理的方法，常见的方法有支持向量机、逻辑回归分类、决策树、随机森林等. 此外，关联规则学习、主成分分析等无监督的算法也被广泛应用于特征基因的筛选，这些方法不仅能够有效地处理大量的基因表达数据，而且能够更加准确地筛选出与癌症相关的特征基因.

在基因表达数据分析的领域，有许多方法用于特征基因的选择．1999年，Golub博士[10]提出了一种以信噪比为基础的特征选择方法，并将该方法应用于人类急性白血病的分类问题，该方法使用加权投票来判断癌症类别，并成功地在7129个特征基因中筛选出了50个与癌症类别高度相关的基因，从而实现了准确分类．其他研究者也提出了许多特征选择方法，如Goh L等人[11]将Person相关系数与信噪比方法相结合，选择出了9个特征基因，其预测精度达到100%；Wang Y等[12]将聚类分析与基因评估排序相结合，提出了新混合特征基因选择算法，选出的基因低冗余且分类准确率较高；Lin等人[13]将遗传算法的适应度函数设置为SVM分类器的准确率，指导特征选择，采用随机采样的策略，只选取23个基因，就可以达到几乎全部分类正确；Gonzalez等人[14]利用模拟退火算法来选择特征子集，只提取了数十个特征基因，就达到了96%的分类准确率；Chuang等人[15]提出了一种以遗传算法和关联度相结合的特征选择算法，并将留一交叉验证的 $k$ 近邻方法作为分类器，从而获得了较高的分类精度；王树林等[16]利用广度优先的搜索算法来构造一个特征基因子集，仅需个位数的基因就可以实现百分百的交叉验证率；陆燕等人[17]在Prim最小生成树的聚类中，采用不同的距离衡量方式，对其进行动态地选取，并使用SVM分类器对其进行预测，从而有效地克服了在高维、非线性数据空间中，分辨率性能降低的问题；关健等[18]提出了一种新方法，首先通过对基因的判别熵值进行统计，来对其在各分类中的贡献程度进行比较，进而利用优化算法对特征基因子集进行寻优，从而获得具有较好性能的特征子集．

近年来，特征基因筛选方法得到了改进，出现了一些混合方法，将多种方法组合起来以提高特征选择的性能并避免各自的缺点．例如，Pavithra等人[19]运用互信息的方法先进行过滤，之后再使用遗传算法进行选取，得到最有特征子集，并使用决策树分类器对结直肠癌数据集进行分类，最后在十折交叉法中验证，仅用8个特征基因就获得了近90%的分类准确率；Xu等人[20]提出一个新的算法，该算法结合了bootstrrap方法和信号噪声比，最终获得了95%的分类精度，比直接对原始数据样本进行分类的结果要高；Radovic M等[21]结合了基因表达量中隐含的时序信息，给出了一种基于最小冗余最大相关的特征选择算法，采用均含有12023个基因的癌症数据集进行验证，并给出了随所选特征不同，其分类准确率的变化．

通过基因表达数据进行胃癌分类的算法需要先用已知类别的训练集训练分

类器，再将分类器应用到测试集上进行预测．在对胃癌进行分类的过程中，训练分类器是十分重要的一个步骤，其优劣对胃癌的分类精度有很大的影响．自从Gloub等人[10]在1999年首先使用基因表达数据进行了人类急性白血病的分类之后，研究者们已经提出了许多对基因表达数据进行分类的改进算法．例如，Furey等人[22]提出的支持向量机算法在多个癌症数据集上进行了分类实验，分类准确率达到了90%以上；Quinlan等人[23]利用C4.5决策树算法进行交叉验证，取得了近95%的准确率；Narayanan等人[24]使用反向传播神经网络的方法对骨髓瘤基因数据集进行分类，使得分类正确率达到87.5%；Karabatak等人[25]利用关联规则(AR)和神经网络(NN)，提出了一种用于胃癌检测的自动诊断系统，最后的分类准确率为95.6%．但是，由于单一的分类器存在着过拟合问题，模型的泛化性能差，预测精度低等问题，因此选取最优的分类器是当前研究的重要部分．

集成分类器相对于单分类器具有更高的泛化能力和分类精度，因此在机器学习领域得到广泛应用，尤其在基因表达谱数据的分类预测方面，研究人员在实际应用中取得了成功．McPherson等人[26]提出的Adaboost算法使用决策树等多种弱分类器，在癌症基因表达数据集上获得了0.91的ROC值；Melville和Mooney[27]提出的多分类器集成方法在人工选定的训练数据集上实验证明其在性能上比其他算法更优，并且提出了一种基于旋转森林的新的多分类器集成算法；Optizhe和Maclin[28]实验证明Bagging算法对大部分情形都有较好的分类效果，但Boosting算法在一些情况下比Bagging效果更好，但可能会过拟合；胡恩召[29]使用强分类算法SVM作为基分类器，研究结直肠癌基因表达数据的分类问题，取得了不错的效果；邵为希[30]针对心脏病数据集建立Stacking集成模型，并通过10折交叉验证防止过拟合，试验表明组合模型的泛化能力和分类性能都优于个体分类器．以上这些研究都证明了利用分类器集成方法来诊断癌症的可行性和有效性．

## 1.3 本文研究内容

在生物信息学研究领域，多位研究者都通过高效的方法来对基因表达数据进行特征筛选，以获取真实数据中的特征基因．这些特征基因在后续分析中被用于癌症分类、与临床数据结合以及在生物领域中提供有意义的解释．

本文以筛选出与胃癌相关的特征基因为目的，从TCGA数据库中下载了胃癌

的基因表达数据作为研究的初始数据集，进行了三个方面的研究．首先，采用基因表达差异分析方法做初步筛选，随后在使用最小冗余最大相关算法选取出最小冗余且与类别最相关的特征基因地基础上，采用基于参数寻优的SVM-RFECV-PSO算法进行特征筛选，得到了19个特征基因用于建模分析．其次，在筛选过的数据上构建胃癌预测模型，分别使用随机森林、XGBoost和Stacking模型评估筛选出的胃癌特征基因的分类效果，发现筛选的特征基因具有很好的区分度．最后，以19个特征基因为变量进行生存分析，探究特征基因与胃癌患者生存状况之间的关系．本文总结现有研究成果，主要分为以下6个章节：

第1章，引言．首先阐述了胃癌特征基因筛选的背景和重要性，并简要地阐述了胃癌及其诊断方法，以及基于集成学习的胃癌预测模型研究的背景和意义．其次，回顾了国内外有关这方面的研究，归纳了文献中的研究思路和成果．

第2章，相关理论及技术．该部分首先介绍了本文研究所用到的数据来源，包括TCGA数据库以及胃癌数据集的描述．接着简单地阐述了三种常见的特征筛选类型，并对本文所需的个体分类模型决策树和支持向量机进行了详细描述，对三种集成学习方法类型进行了详尽地描述，此外，还给出了模型的评价准则．

第3章，胃癌预测模型的构建与分析．首先采用FDR误差控制法进行初步筛选，得到差异表达基因．接着采用mRMR算法进一步进行筛选，得到300个最小冗余最大相关的差异表达基因．最后，结合了SVM-RFE算法和粒子群算法，提出新的特征选择模型SVM-RFE-PSO，从300个差异表达基因中筛选出19个胃癌特征基因，为后续的集成学习算法提供基础．

第4章，集成学习的构建与结果分析．基于SVM-RFECV-PSO算法特征选择的基础上，使用基于Bagging的随机森林算法和基于Boosting的XGBoost算法进行建模，实验得到正确率和AUC值分别为95.12%、94.31%和0.9923、0.9867．然后，使用结合多种算法的Stacking模型对是否患胃癌进行预测，得到98.52%的准确率以及0.9979的AUC值，说明筛选的特征基因对是否患癌癌的分类具有很好的区分度，筛选方法可靠．从预测效果看，Stacking模型对胃癌发病的预测效果最好．

第5章，胃癌患者的临床数据分析．在本章中，我们探讨了如何将特征基因和临床数据结合来进行深入分析，介绍了生存分析的原理，讨论了如何利用K-M模型进行生存分析，并解释了特征基因的组学信息．

第6章，总结与展望．本章总结了已完成的工作，并对本论文的局限性进行了分析和讨论，同时展望了未来研究的方向．

# 第2章 相关理论基础及介绍

## 2.1 数据来源及描述

为了进行后续的特征选择和分析，获取完整的胃癌基因表达数据集是非常关键和必要的．在本节中，我们介绍了一个可靠的数据库——TCGA数据库，可以用来下载胃癌数据集．此外，我们还讲解了如何使用R语言中的TCGABiolinks包来下载胃癌数据，这是进行后续分析的重要基础步骤．

### 2.1.1 TCGA数据库介绍

TCGA是“The Cancer Genome Atlas”的缩写，该项目始于2005年，旨在通过基因序列分析和生物信息学方法对肿瘤相关基因突变进行编目．通过高通量基因组分析的技术，TCGA能够帮助研究者更好地了解不同癌症类型的遗传学特征和机制，包括癌症的基因突变、基因表达谱、DNA甲基化、基因拷贝数变异等．该项目受美国癌症研究所下属的癌症基因组中心和美国人类基因组研究所的监管．

TCGA项目由两个主要部分组成：基因组表征中心(genome characterization centers, GCCs)和基因组数据分析中心(genome data analysis centers, GDACs)．其中GCCs主要负责进行测序，GDACs则负责对测序数据进行分析．截至目前，TCGA已经收集了39种癌症相关的测序数据，覆盖了29个癌症器官，包含1万多个肿瘤样本和27万多个文件．TCGA的数据类型主要包括突变(Mutation)、拷贝数(Copy Number)、mRNA、miRNA、DNA甲基化(Methylation)、临床信息(Clinical)和蛋白质(Protein))等．可以从表2-1看到TCGA数据库包含的数据类型[31]．

表2-1 TCGA数据库中包含的数据类别

| 类别 | 详细信息 |
|---|---|
| mRNA | 肿瘤样本的基因表达情况 |
| miRNA | 肿瘤样本中微小RNA的表达情况 |
| Methylation | 肿瘤样本DNA的甲基化情况 |