单位代码 10635

学　　号

西南大學

硕士学位论文

基于深度学习和组学数据融合的乳腺癌生存期预测研究

论文作者：

指导教师：

学科专业：计算机系统结构

研究方向：生物信息学

提交论文日期：　　年　　月　　日

论文答辩日期：　　年　　月　　日

学位授予单位：　西南大学

中 国 • 重 庆

20　　年　　月

独创性声明

学位论文题目：基于深度学习和组学数据融合的乳腺癌生存期预测研究

本人提交的学位论文是在导师指导下进行的研究工作及取得的研究成果。论文中引用他人已经发表或出版过的研究成果，文中已加了特别标注。对本研究及学位论文撰写曾做出贡献的老师、朋友、同仁在文中作了明确说明并表示衷心感谢。

学位论文作者：　　签字日期：　　年　　月　　日

学位论文版权使用授权书

本学位论文作者完全了解西南大学有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权西南大学研究生院可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

本论文公开时间：□获学位当年；　□推迟 1 年。

学位论文作者签名：　　导师签名：

签字日期：　　年　　月　　日　　签字日期：　　年　　月　　日

目　录

基于深度学习和组学数据融合的乳腺癌生存期预测研究

计算机系统结构专业 硕士研究生 XXX
指导老师 XXX

摘 要

近年来，乳腺癌的发病率和死亡率持续升高，对乳腺癌患者进行精准的生存期预测已成为癌症研究领域的热点问题。准确的生存期预测能够为医务工作者和病人家属提供科学的治疗凭据，同时避免患者过度治疗所造成的医疗资源浪费。

乳腺癌是一种恶性肿瘤疾病，它的产生和发展与基因密切相关。随着生物测序技术的进步，生物信息学领域积累了大规模组学数据，这为研究者全方位了解生物学过程夯实了基础。在乳腺癌生存期预测研究中，基因表达数据从微观生物学层面反映肿瘤的生物特性，对癌症预后和治疗有重要的应用价值。临床数据包含了丰富的病理学特征，为乳腺癌患者的生存期预测提供理论依据。如何有效地融合基因表达数据和临床数据，更准确地预测乳腺癌生存期，是癌症生存期预测研究领域中亟需解决的问题。然而，现有的乳腺癌生存期预测模型，往往使用单一的特征选择方法对基因表达数据进行特征提取，再进行简单的特征拼接融合，这不仅容易丢失重要的基因信息，还忽略了组学数据间的关联。因此这类方法具有一定的局限性。

本文在现有乳腺癌生存期预测研究的基础上，提出了基于深度学习和组学数据融合的乳腺癌生存期预测模型。首先，通过非负矩阵分解的改进算法（Multi_NMF）提取出与乳腺癌生存期相关的特征基因；其次，构建基于 Attention 机制的深度神经网络模型（AMND）来融合基因表达数据和临床数据；最后，在上述研究的基础上，构建了一种基于多尺度特征融合的生存期预测模型（MFFD）来获取不同尺度上的组学特征。实验结果表明，本文的方法相较于现有的乳腺癌生存期预测方法，具有更好的预测效果。本文主要完成以下三个方面的工作：

（1）本文在非负矩阵分解（Nonnegative Matrix Factorization，NMF）算法的基础上，提出了一种基于非负矩阵分解算法的 Multi_NMF 特征选择算法。Multi_NMF 方法不仅可以提取基因表达数据的高阶特征，还避免了由矩阵分解所

带来的稀疏性和丢失重要特征等问题。实验结果表明，改进后的 Multi_NMF 方法能够选择出更多与乳腺癌生存期相关的基因，有利于提升乳腺癌生存期预测准确率。

（2）为了获取更多的组学数据特征，本文提出了一种基于组学数据和 Attention 机制的深度神经网络模型（AMND）来融合患者的基因表达数据和临床数据。作为 Attention 机制在乳腺癌生存期预测模型上的初步尝试，AMND 方法考虑了不同数据间的差异性，通过计算临床数据和基因表达数据的相关性，自适应地融合来自不同特征提取方法的特征基因，从而提升乳腺癌生存期预测准确率。实验结果表明，AMND 方法能够准确地预测出乳腺癌患者的生存期。

（3）针对数据集样本数量小而导致模型不能有效学习的问题，本文提出了基于多尺度特征融合的生存期预测模型（MFFD）。首先，使用 AMND 方法融合组学数据。其次，设计不同的池化层以获取不同尺度的特征。最后，将不同尺度的特征拼接融合。MFFD 方法融合了不同粒度的组学数据特征，包含了更多的生物特征信息。通过在测试集上的性能评估，实验结果表明，MFFD 方法进一步提升了乳腺癌生存期预测性能。

关键词：乳腺癌；非负矩阵分解；注意力机制；组学数据；多尺度

Research on Breast Cancer Survival Prediction Based on Deep Learning and Omics Data Fusion

Major: Computer Architecture

Master: XXX

Supervisor: XXX

Abstract

Recently, the morbidity and mortality of breast carcinoma are both increasing gradually. As a result, it has become a hot issue in cancer research to make precise prognostic prediction for breast cancer patients, which can help not only to effectively avoid overtreatment and medical resources waste, but also to provide scientific basis to assist medical staff and patients' family members to make right medical decisions.

Breast cancer is a malignant tumor disease, whose emergence and development are closely related to genes. With the advancement of DNA sequencing technique, large-scale omics data has been accumulated in the field of bioinformatics, which paves a solid foundation for researchers to comprehensively understand biological processes. In the study of breast cancer survival prediction, gene expression data reflects the biological characteristics of tumor from the micro biological level, which has important application value for cancer prognosis and treatment. The clinical data contain abundant pathological features, which provide theoretical basis for survival prediction of breast cancer patients. How to effectively integrate gene expression data and clinical data for prognostic prediction of breast cancer is an urgent problem in the field of cancer survival prediction. However, existing breast cancer survival prediction models tend to use a single feature selection method to extract the features of gene expression data, and then conducts simple spicing and fusion of feature data, which can easily lead to the loss of important gene information and the neglect of the correlation between omics data. Therefore, this sort of methods does have limitations.

Based on existing research of breast cancer survival prediction, this paper proposes a model based on deep learning and omics data fusion. Firstly, an improved nonnegative matrix factorization algorithm (Multi_NMF) is proposed to extract the feature genes related to breast cancer survival. Then, a deep neural network based on attention mechanism (AMND) is constructed to fuse gene expression data and clinical data. Finally, on the basis of above researches, a deep neural network model based on multi-scale feature fusion (MFFD) is introduced. The experimental results show that compared with existing methods, the method in this paper has better prediction performance. The main research contents of this article are summarized as follows:

(1) Based on the nonnegative matrix factorization (NMF) algorithm, Multi_NMF feature selection algorithm is proposed. This method can not only extract high-level features of gene expression data, but also avoid the problems of sparsity and loss of important feature information caused by matrix decomposition. The experimental results show that the improved Multi_NMF method can select more informative genes related to breast cancer prognosis and thus more accurate prediction can be obtained.

(2) In order to explore the effectiveness of omics data for the breast cancer survival prediction, this paper proposes a deep neural network model based on omics data and attention mechanism (AMND) to combine gene expression data and clinical data. As an initial attempt of attention mechanism in the breast cancer prognosis, AMND method is able to better consider the connection between clinical data and gene expression data and self-adaptively fuse of feature genes from different feature extraction methods, improving the accuracy of breast cancer survival prediction. The experimental results show that AMND method can accurately predict the survival time of breast cancer patients.

(3) In order to solve the problem that the model cannot effectively study due to the small number of samples, this paper proposes a deep neural network model based on multi-scale feature fusion (MFFD). It combines the features of different granularity of the group data, and contains more feature information. Through performance evaluation on the test set, experimental results show that MFFD further improves breast cancer survival prediction performance.

Keywords: breast cancer; nonnegative matrix decomposition; attention mechanism; omics data; multi-scale

第 1 章 引言

1.1 研究背景及意义

乳腺癌作为女性最常见的恶性肿瘤[1-4]，已成为世界性的重大健康问题。据世界癌症统计报告显示[5]，全球每年约有 120 万新增乳腺癌患者，其中约有 50 万女性死于乳腺癌。乳腺癌是一种全身性疾病，经临床治疗后将近一半的患者可治愈，另一半的患者可能出现复发或转移的状况。尽早检测出经临床治疗后仍未痊愈的患者，并对其进行辅助治疗，将有效提高乳腺癌患者的生存率。乳腺癌患者的生存期受多种因素影响，如年龄、肿瘤大小、肿瘤的组织学分级、诊断时局部及远处转移的程度、激素受体状态、HER-2 表达情况、患者的治疗时间和对治疗的依从性等。研究指出[6]每个基因都有其特定的功能，可以对患者的生理状况产生不同程度的影响。因此，基于多基因的研究有助于科研工作者详细观察癌细胞转移情况，从而进一步改善乳腺癌生存期预测效果。

组学数据[7]来源于高通量实验技术产生的海量数据。例如，全基因组检测技术检测到的转录组和表观组等多个层面的数据以及基于微阵列技术所得到的数据等。研究中常用的组学数据有 DNA 甲基化、miRNA 表达数据、基因表达数据、蛋白质表达数据和拷贝数变异等。乳腺癌的预后与基因标志物密切相关，而组学数据细致描述了肿瘤的分子机制，为人们洞悉生物系统奠定了良好的理论基础[8]。同时，组学数据与肿瘤癌变的生物过程有紧密联系，如 DNA 甲基化和 miRNA 表达数据均可以影响基因表达水平[9]。因此，使用组学数据来进行生存期预测已经成为乳腺癌生存期预测研究的热点问题。

随着人们对深度学习和机器学习的日益关注，越来越多的机器学习算法和深度学习技术用于图像处理、文本分类等领域，并取得显著效果。例如，研究人员使用机器学习算法，对高维度基因表达数据建立有效的聚类模型。非负矩阵分解算法作为机器学习中常用的方法。近年来国内外学者对其优化，得到一系列改进的非负矩阵分解算法。该算法能对高维度的基因表达数据进行降维和局部特征提取。因此，利用这些技术，我们可以开发高效的计算方法来精确地预测乳腺癌患者的生存期。这不仅可以帮助乳腺癌患者了解其预期寿命，还可以帮助临床医生做出科学的决策并进一步指导后续治疗。同时，降低了乳腺癌的总体死亡率，提高了乳腺癌患者的生活质量，具有十分重要的研究意义。

综上所述，融合组学数据已成为乳腺癌生存期预测研究的关键，如何构建精准高效的组学数据融合模型成为研究热点。将深度学习领域中的 Attention 机制和多尺度特征融合技术应用于乳腺癌症生存期预测研究，是一个非常有意义的研究

课题。

1.2 国内外研究现状

癌症生存期预测研究初期，科研人员主要从基因表达数据中提取重要的特征。例如，Van't Veer 等人[10]采用多变量分析方法，在 98 例乳腺癌病人的基因表达数据中，确定了 70 个与淋巴结阴性乳腺癌预后相关的基因。Glas 等人[11]使用高通量技术将 70 个基因表达谱转译为 MammaPrint 基因芯片，并将其作为年轻乳腺癌患者预后预测的重要指标。上述研究证实了乳腺癌生存期预测与基因标志物密切相关，然而使用传统的统计学方法选取基因标志物，存在信息不全面等问题。Xu 等人[12]使用基于支持向量机的递归特征消除（SVM-RFE）法得到有利于生存期预测的重要特征。这些特征可以有效地预测细胞转移，从而将预后不佳的患者区分开来。由此可见，相较于早期的特征筛选方法，使用机器学习算法进行特征提取能获得更好的分类效果。

随着新一代测序技术的发展，其他类型的基因组数据也应用于研究。如 DNA 甲基化[13]、miRNA[14]和拷贝数变异（CNV）[15]。然而，每种数据源所表现的基因功能是不一样的，难以获得准确的生存期预测结果。为了在基因组学层面上对病人有更全面的认知，除了改进基因表达数据的特征提取方法，研究人员发现融合组学数据可以获得更好的预测性能。例如，Marietteet 等人[16]使用无监督的多核框架预测乳腺癌的临床结果。Kim 等人[17]设计了一个基于语法的神经网络来评估卵巢癌的预后。Ahmad 等人[18]提出了一种将高斯混合模型与加速失效时间模型相结合的分层贝叶斯模型，来发现与乳腺癌临床相关的疾病亚型。Coretto 等人[19]通过结合稀疏关系矩阵估计器和最大似然估计法来识别不同癌症风险亚型的差异表达基因。然而，这些方法局限于从成千上万的数据样本中获取特征，未充分考虑不同组学数据间的联系。

近几年来，深度学习技术在生物信息学[20]等诸多领域被广泛应用，并取得较好的效果。例如，Chaudhary 等人[21]通过无监督自编码模型整合 RNA-seq、miRNA-seq 和 DNA 甲基化数据，重建具有代表性的复合特征。然而，该研究采用的无监督分类算法只能将患者分为两组，并不能直接将复合特征与生存时间联系起来，使得预测生存期的准确度有限。Chai 等人[22]提出了一个名为 DCAP 的深度学习模型，该模型使用自动编码网络来提取多组学数据中的重要特征，然后将这些特征输入到比例风险模型中，来预测癌症的预后。实验结果表明，与单一 mRNA 数据相比，基于组学数据的模型比以前的方法至少提高了 7.4%的精确度。Gevaert 等人[23]使用基于贝叶斯网络的概率模型将乳腺癌病人的临床信息和基因表达数据融合，在 METABRIC 数据集上，5 年生存期预测准确率达到 82%。Sun 等人[24]在

I-RELIEF 算法的基础上提出了一种新的数学模型来预测乳腺癌转移的可能性。该方法综合了临床信息中包含的生物特征和基因表达数据中的遗传物质，实验结果表明混合特征可以显著提高预后的特异性。上述研究表明，考虑多种数据能有效提高乳腺癌生存期预测的准确性。但是，上述工作的特征选择方法过于单一，对组学数据的特征提取不够全面。这不仅容易丢失重要的基因信息，还忽略了组学数据间的关联，依然具有局限性。

最近，基于 Attention 机制的神经网络在各个领域成为研究热点，如 Bahdanau 等人[25]在 encode-decode 框架的基础上，实现了机器翻译中的对齐与翻译功能，该模型的性能在英法语数据的评测上超过了使用基于短语的机器翻译系统。注意力机制被 Luong[26]等人用来对单个词进行权重分配，从而实现语句的精准翻译。Google Mind 团队[27]将 Attention 机制应用在图像分类上，该方法能够捕捉图像的局部特征，有效提高了分类准确率。Zhang 等人[28]采用多信息融合的方式，提出了一种基于 Attention 机制的深度学习模型，该模型可以自适应的学习上下文信息中的最优特征。多尺度特征融合技术是常见的像素级融合方法[29]，该方法对不同粒度、不同结构特征的细节信息进行处理，可以获取比较好的融合效果。

1.3 本文主要研究内容

基于上述研究背景和课题意义，本文主要研究内容如下：

（1）乳腺癌基因表达数据的维度较高，其中包含了大量的无关特征，这将严重影响模型的效率和预测结果。为了降低基因表达数据的维度，获取与乳腺癌生存期相关的重要特征，本文采用改进的非负矩阵分解算法（Multi_NMF）对基因表达数据进行特征提取。通过 Multi_NMF 算法提取的特征，包含了更全面的基因信息，具有更好的预测效果。

（2）不同的数据包含不同的特征信息。针对单一数据无法全面表达患者信息的问题，本文提出一种基于 Attention 机制的深度神经网络模型来融合组学数据。首先，使用 Multi_NMF 方法获取基因表达数据的特征。其次，考虑到不同组学数据的相关性，引入 Attention 机制，根据临床数据分别计算不同 NMF 优化算法得到的特征矩阵的权重，并对特征向量进行加权求和。此时，得到新的基因表达特征向量，再融合该样本的临床数据，并将其输入深度神经网络进行预测。该方法不仅考虑了组学数据，还自适应地融合了多种特征提取方法，能充分提取基因表达数据和临床数据的高阶特征，从而提升乳腺癌生存期预测性能。

（3）针对固定感受野获得的组学数据特征具有尺度不变性的问题，本文提出了基于多尺度特征融合的生存期预测模型（MFFD）。首先使用第四章提出的方法融合组学数据。其次，设计多尺度特征融合的神经网络，将组学数据输入不同大

小的池化层进一步提取不同粒度的特征信息。最后将这些特征拼接融合，得到组学数据的深层次特征。该方法不仅提高了重要特征的利用率，还优化了乳腺癌生存期预测模型的性能。

本文主要研究内容和文章的整体结构如图 1.1 所示：

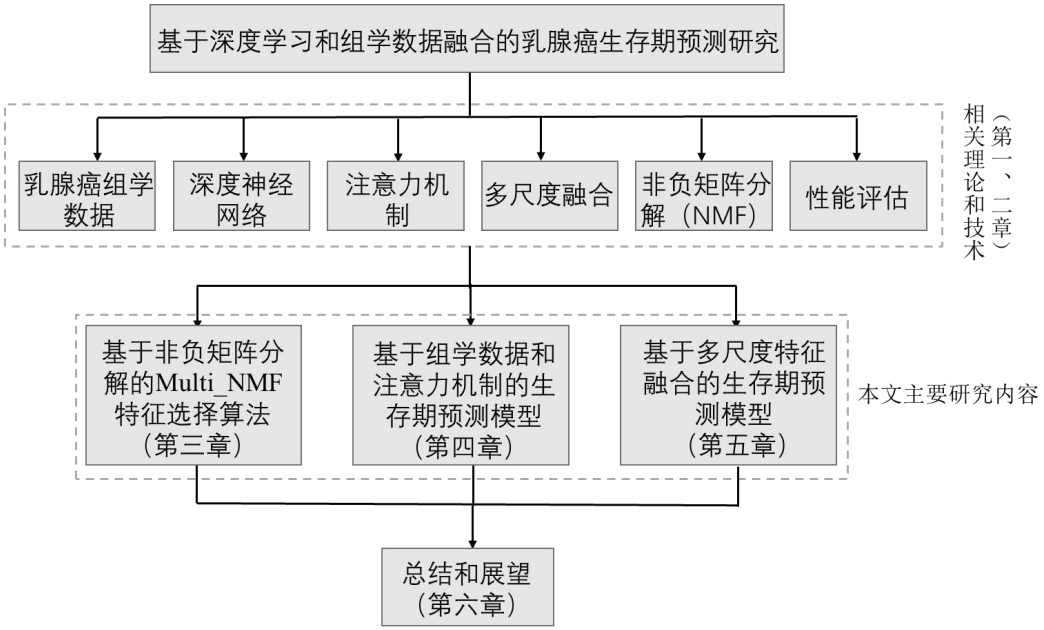

图 1.1　本文研究内容框架图

1.4 论文结构

文章的整体结构和每章的内容如下所述：

第 1 章，引言。首先，本文概述了乳腺癌生存期预测的研究背景和意义。其次，对癌症生存期预测的国内外研究现状和应用领域进行了叙述。最后，介绍了文章的主要研究内容和论文结构。

第 2 章，相关理论与技术。重点阐述文章涉及的网络模型、算法和相关理论。首先描述乳腺癌组学数据的相关理论。其次，对深度学习的相关理论进行概述，如常用的深度神经网络模型和注意力机制等。然后介绍了非负矩阵分解算法的相关理论。最后，介绍了本文进行乳腺癌生存期预测研究所使用的评价指标。

第 3 章，基于非负矩阵分解算法改进的 Multi_NMF 特征提取方法。针对单一非负矩阵分解算法无法全面得到重要特征的缺陷，提出一种改进的 Multi_NMF 算法。并且，在 METABRIC 乳腺癌数据集和 TCGA 乳腺癌数据集上对 Multi_NMF 算法的特征选择性能进行比较与分析。

第 4 章，基于组学数据和注意力机制的生存期预测模型。针对当前乳腺癌生存期预测模型存在未考虑组学数据间的联系和基因高阶特征表示不完全的问题，本章引入 Attention 机制，设计了基于 Attention 机制的生存期预测模型。该模型通

过计算临床数据和基因表达数据的相似度，获得具有区分能力的加权因子，得到更具有表示能力的基因特征。最后，通过设置多角度的对比实验来测试该模型的性能。

第 5 章，基于多尺度特征融合的生存期预测模型。首先，本文对第四章提出的基于组学数据和注意力机制的生存期预测模型进一步优化，得到基于多尺度特征融合的深度神经网络模型（MFFD）。该模型能从多尺度获取组学数据的特征，更好的表示患者的详细信息。其次，描述 MFFD 模型的结构和网络参数。最后，设计多角度的对比实验来验证 MFFD 模型的有效性。

第 6 章，总结与展望。对本文现阶段所做的工作和研究成果进行总结，分析及讨论模型中的不足之处，并对未来要做的工作进行展望。

1.5 本章小结

本章首先概述了乳腺癌生存期预测的研究背景和意义。其次，对国内外乳腺癌生存期预测研究的现状进行了总结和说明。最后，介绍了本文的主要研究内容和文章的整体结构。

第 2 章 相关理论与技术

2.1 乳腺癌组学数据介绍

近年来，DNA 检测技术发展迅速，越来越多与癌症相关的重要基因被发现，这使得癌症数据库在扩大的同时也得到了进一步的完善。这些数据中包含了患者的详细信息，如临床信息、病理图像数据和组学数据，它们为科研工作者的研究提供了数据基础。因此，融合多种数据对乳腺癌患者进行生存期预测是癌症研究领域的重点。不同的数据集有不同的特点，其中包含的癌症类型也略有差异。有些数据集记录了同一类型的癌症相关数据，而其他数据集则包含不同癌症类型的组学数据。例如，METABRIC[30]数据集记录了与乳腺癌相关的组学数据和临床数据。TCGA[31]、NKI[32]、SEER[33]和 TCIA[34]等数据库收录了乳腺癌、直肠癌、胃癌、卵巢癌等多类癌症的组学数据。这些数据集能极大地促进人们对乳腺癌生存期的预测、分析和可视化研究，为生物医学的发展奠定了基础。

本文使用的组学数据来自 METABRIC 数据集和 TCGA 数据库。TCGA（The Cancer Genome Atlas，癌症基因组图谱）是当前最大的癌症基因组数据库，该数据库收录了 20 多种癌症的组学数据，如基因表达数据、临床数据和 DNA 甲基化等，是癌症研究的重要数据来源。METABRIC（Molecular Taxonomy of Breast Cancer International Consortium）是英国和加拿大合作的医疗项目，旨在根据细胞分子特征将乳腺癌进一步细分为更多的分子分型。例如，浸润性乳腺癌可以分为四种分子亚型，分别是 HER2 过表达型、官腔上皮 A 型、Luminal B 型和三阴型。同时，METABRIC 是一种具有较长生存期的乳腺癌数据集[35]。METABRIC 数据集中包含了 1980 例乳腺癌病人的临床信息和诸如拷贝数变异、基因表达数据、SPN 等组学数据。本文建立了两个乳腺癌组学数据集，其中，METABRIC 数据集用于初步测试模型的基本功能，TCGA 数据集作为独立测试集用来进一步验证模型。以下分别介绍组学数据中的基因表达数据和临床数据。

2.1.1 基因表达数据

受益于生物信息学和微阵列技术的普遍应用，科研工作者对肿瘤的认识已从细胞水平深入到基因水平。因此，相关研究者开始从基因层面对癌症进行诊断和预测。基因表达数据[36]是微阵列技术在芯片上测量不同样本的基因在不同状态和不同组织中的 mRNA 水平得到的一组数据，它展示了基因转录产物在细胞中的丰度以及基因调控信息。基因表达数据具有小样本、高维度和高噪声的特点，且存在大量冗余信息。小样本数量让许多基于大样本的统计学方法和机器学习方法无法在该数据集上获得有统计意义的结果[37]。从机器学习角度来看，远超过样本数

量的高维度数据容易遭遇维度灾难[38]。事实上，只有部分基因与乳腺癌生存期预测紧密相关，我们将这些基因称为特征基因。因此需要使用更好的特征选择算法剔除冗余基因，降低数据的维度，从而获取高阶表达的特征。

本文使用 TCGA 数据集和 METABRIC 数据集中的基因表达数据。以 TCGA 数据集为例，该数据集包含了 17877 个基因的表达水平。表 2.1 展示了部分患者的基因表达数据。从表 2.1 中可知，第一列是基因名，从第二列起，每列为一个样本病例，矩阵中的数据为对应患者的基因表达值。例如，TCGA-AC-A3W5-01 样本在 GCLC 基因上的表达值为 3113。

表 2.1　TCGA 基因表达数据格式

Gene	TCGA-AC-A3W5-01	TCGA-BH-A8FZ-01	TCGA-AC-A23G-01
GCLC	3113	1894	1359
TNMD	27	48	7
DPM1	1608	1828	623
SCYL3	2371	1049	891
FGR	1242	1935	473

2.1.2 临床数据

临床医学数据有时间序列数据、截面数据和纵向数据（Longitudinal data）等形式。截面数据是指在某一时刻对不同对象进行调查得到的数据，在临床上可以用来分析疾病的影响因素[39]。随着时间变化收集到的数据是时间序列数据。纵向数据是指患者在随访期间的数据[40]，它不仅包含截面数据的特点，还具有时间序列数据的特点。临床医学纵向数据，包含了病人个体特征随时间变化的趋势，也蕴含着个体之间的差异性，有着很重要的医学分析价值。现有研究表明[41]，临床数据与癌症生存期预测研究密切相关。

表 2.2　METABRIC 临床数据格式

Patient_ID	OS_Months	Age	Vital_Status	ER_IHC
MB-0000	140.5	75	Living	Pos
MB-0002	84.6	43	Living	Pos
MB-0005	163.1	48	Died	Pos
MB-0006	164.9	47	Living	Pos
MB-0008	41.3	76	Died	Pos

本文使用 TCGA 数据集和 METABRIC 数据集中的临床数据。以 METABRIC 数据集为例，每个病例有 27 维临床特征，如诊断年龄、组织学分级、肿瘤大小、细胞转移状态以及病理分期等。表 2.2 展示了 METABRIC 数据集的部分临床数据。其中，第一列表示样本的 ID 号，后面依次是该病人的临床特征，如总生存时间、诊断年龄、存活状态等特征。

2.2 深度学习相关理论

深度学习能够模拟人脑神经系统构建深层神经网络模型，通过无监督或者有监督的方式从大量数据中学习丰富的层级特征，这些特征可用于模式识别和分类任务。深度学习模型的优点有以下几个方面：（1）深度学习模型有很多隐藏层，这样的特征空间可以映射到任意函数，从而解决现实生活中的复杂问题。（2）基于大数据的深度学习模型具备更优越的建模能力，数据量越大，模型的表现越好。（3）深度学习模型的学习能力非常强，可以通过网络学习到数据中的重要特征，从而达到准确的分类或预测。（4）模型具有较强的可移植性。目前有很多深度学习框架的表现优异且兼容多个平台，如 TensorFlow 和 Pytorch。经过多年的努力，科研工作者不仅在理论上取得了丰硕的学术成果，同时在图像处理、语音识别以及文本挖掘等多个应用场景也取得了较好的效果。下面将简要介绍深度学习中主流的网络模型和相关技术。

2.2.1 深度神经网络

作为常用的深度学习框架，深度神经网络（Deep Neural Networks，DNN），可看作一个有很多隐藏层的感知机。其中，相邻两层网络节点的连接方式为全连接。深度神经网络相较于浅层神经网络在建模方面更具有优势，它能够有效处理复杂的非线性系统，并且利用网络中的隐藏层对数据进行高层次抽象，使整个模型的泛化能力得到提升。按不同层的位置划分，DNN 内部的神经网络层可以分为：输入层、隐藏层和输出层。其结构如图 2.1 所示。

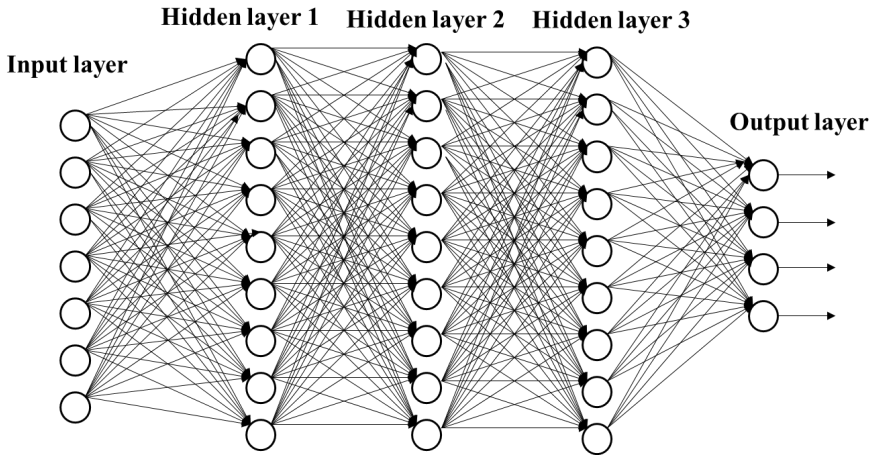

图 2.1　深度神经网络示意图

其中，数据通过输入层传入模型，模型内部则由多个隐藏层构成，每个隐藏层包含多个神经元。隐藏层的主要作用是处理上一层传递过来的数据信息，并且将信息传递到下一层。模型的最后一层为输出层，输出层的维度由任务的特性决定。在感知机中，各层的神经元通过全连接的方式组织在一起，将信息由前往后

传递。对前一层神经元的数值进行加权求和就可以得到后一层神经元的值。计算方法如公式（2-1）所示。

$$h_j = \sigma\left(\sum_i w_i x_i + b_i\right) \tag{2-1}$$

在公式（2-1）中，w 代表权重参数，b 表示函数偏置值，σ 表示非线性激活函数。激活函数使神经网络具有非线性映射能力，通常采用 *Relu*，*tanh* 和 *sigmoid* 等函数表示，其数学表达如公式（2-2）到（2-4）所示。

$$sigmoid(x) = \frac{1}{1+e^{-x}} \tag{2-2}$$

$$tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{2-3}$$

$$Relu(x) = \begin{cases} x, & x > 0 \\ 0, & x \le 0 \end{cases} \tag{2-4}$$

在激活函数的作用下，神经网络近似等同于非线性函数，并且通过神经网络训练出的非线性模型能够应用在预测、分类等任务中。激活函数的选择需依据实际情况并结合不同激活函数的特点进行配置。

神经网络的学习过程主要包括两个方面：前向传播（Forward Propagation，FP）和反向传播（Backward Propagation，BP）[42]。数据通过输入层进入网络，经过逐层计算得到最终的输出结果，该过程称之为前向传播。根据损失函数计算输出值与输入数据的残差，采用链式求导法则计算出每个权值的梯度，并由反向传播算法更新每个神经元的权值，从而优化网络参数，该过程为反向传播[43]。反向传播过程中，使用随机梯度下降（Stochastic Gradient Descent，SGD）方法[44]对网络参数进行调整。而深度神经网络中采用的参数优化方式是小批量随机梯度下降算法[45]。小批量随机梯度下降方法可以通过矩阵计算和向量计算等方式来提升运行速度，还可以使用降低迭代次数的方法使收敛更加稳定。

2.2.2 卷积神经网络

卷积神经网络（Convolutional Neural Networks，CNN）作为深度学习中常用的网络结构之一[46,47]，是一种通过卷积计算对输入数据进行处理的前馈神经网络。卷积神经网络的应用范围十分广泛，在语音识别、图像处理等领域做出了突出贡献。其网络结构与生物神经网络类似，通过权值共享有效减少了深度学习网络模型的计算复杂度。对于网格结构数据[48]，卷积神经网络的优势更为突出。例如，图像作为典型的网格数据可以直接输入卷积神经网络，规避了复杂的图像数据处理过程，如特征提取和数据重建。另外，卷积神经网络对数据进行比例缩放、平移、倾斜以及其他形式转化时，能够保持数据的高度一致性。图 2.2 展示了一个简

化的卷积神经网络架构图。

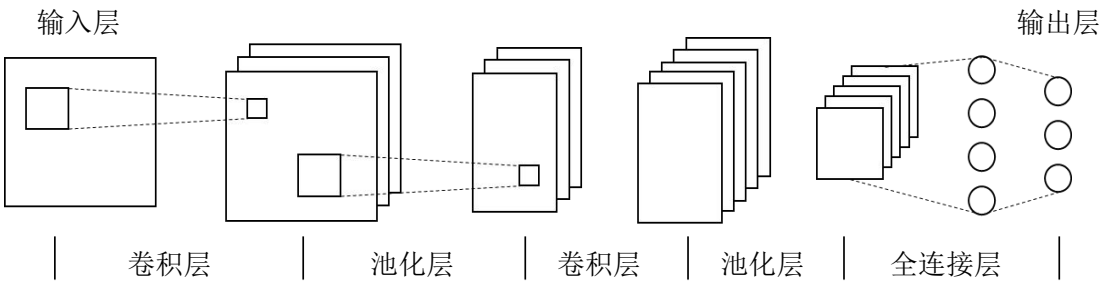

图 2.2　卷积神经网络架构图

从图 2.2 中可以看出，卷积神经网络主要由若干个卷积层、池化层和全连接层组成。其中，前几个网络层主要是以卷积层和池化层为主，特征图将卷积层中的各个单元组织在一起。特征图中的每一个计算单元都通过滤波器连接上一个特征图中对应的局部块，并将局部块中的权值进行加权求和，所得结果传递给非线性函数。该函数一般采用激活函数，如 *sigmoid* 函数。同一张特征图上的神经元使用的过滤器都是相同的，而其他各层特征图所采用的过滤器略有差异。这样的处理方式对于多维数据更具有优势。在多维数据中，每一个像素点与其周围像素点的值互相关联，这样能够得到图像中重要的局部特征。此外，其他位置的局部特征之间略有差异，使得不同位置的神经元均可以对权值进行共享。

卷积层的主要作用是对局部特征进行提取，池化层的作用是对相近的特征进行整合。通常，池化层采用最大池化或者平均池化对特征图中的局部块进行处理，在读取数据时通过移动一行或一列的方式进行操作，这样的方式能够有效降低输入数据的维度以及保证数据的移动不变性。通过卷积层、池化层、非线性变换的计算和最后的全连接操作，神经网络能够直接对输入数据进行特征提取。如果是处理分类任务，可以在最后加入一个分类层，这样就实现了一个完整的神经网络分类器。

2.2.3 注意力机制介绍

注意力（Attention）机制是深度学习领域的一种新技术。它的主要思想是度量键（Key）和请求（Query）之间的相似度，若某一时刻的输入与目标状态越相似，那么这一时刻的输入所对应的权重就会越大，说明当前的输出更依赖于该时刻的输入。例如，我们翻译“机器学习”这一词语。当翻译成“machine learning”的时候，我们希望模型更加关注的是“machine”而不是“learning”。如图 2.3 所示，Attention 体现为 h^1 和 z^0 的匹配度。

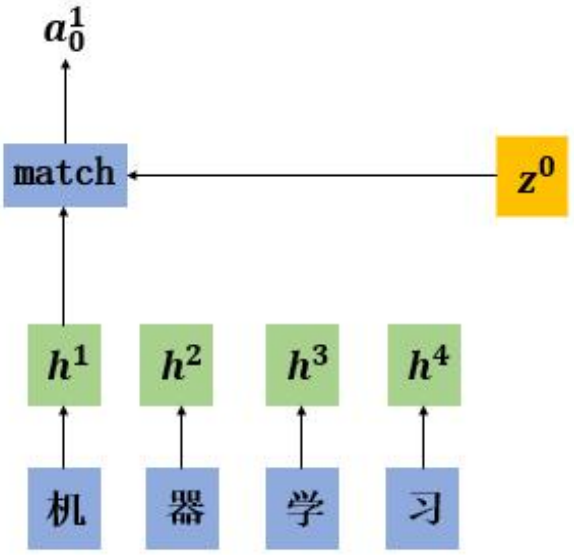

图 2.3 “机器学习”翻译示意图

其中，h^1 为网络隐藏层的输出向量，z^0 为初始化向量，match 为计算这两个向量匹配度的模块，a_0^1 是 match 算出来的相似度。相似度的计算方法有很多，如余弦相似度、神经网络和矩阵变换等方法。Attention 的最终目标是得到基于所有时刻输入的加权和，所以我们期望权值之和为 1。在这里，我们使用 *softmax* 函数达到该目标。最终的加权向量和 c^0，作为神经网络的输入，并通过该输入得到最终的预测结果。如图 2.4 所示。

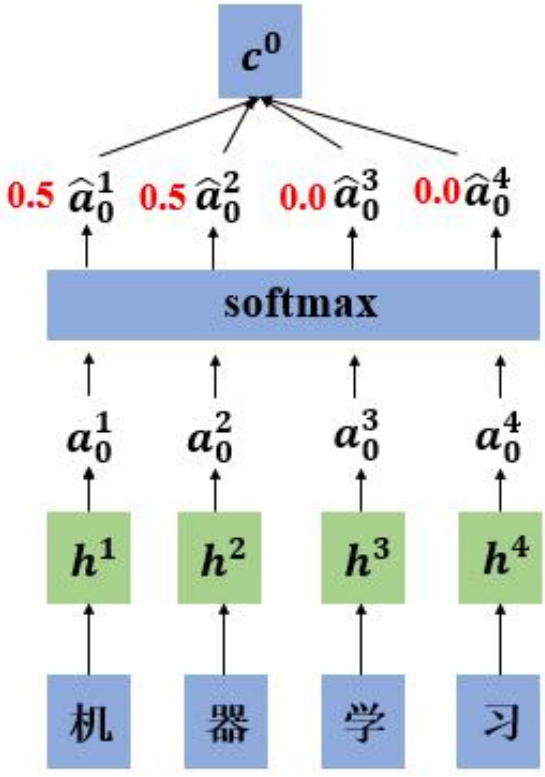

图 2.4 Attention 机制示意图

2.2.4 多尺度模型介绍

所谓多尺度，就是对信号不同粒度的采样。事实上，物体的大量特征存在于不同的空间上，人们从不同的角度去观察物体，会看到物体具有不同的表现形式。例如，我们远距离观察物体，可以看到物体的几何外形、轮廓等宏观特征，当我们在近处观察物体，可以看见物体的内部组成、表面纹理等细节特征。所以，采取恰当的尺度去观察和分析目标，从而获取更全面的特征信息具有重要的研究意义。

感受野作为计算机视觉领域中一个非常重要的概念，用来表示深度神经网络中各个神经元对原始目标的感受范围。深度神经网络通过卷积层和池化层对局部区域进行特征提取，使得神经元无法对原始目标进行全面感知。感受野能接触到的原始数据的范围与感受野的值成正相关，感受野的值越大意味着神经元提取到的特征更全面、语义层次更高；感受野的值越小则意味着提取到的特征越细致。因此研究人员一直在设计各种各样的多尺度模型。

多尺度特征融合模型本质上是在不同的感受野下进行特征提取，再进行拼接融合。例如，在 Inception 网络中，将 1×1、3×3 和 5×5 等维度的卷积并列放置在一起构成了 Inception 基本模块，如图 2.5(a)所示。Inception 模块能够适当的增加网络的宽度和深度，来提高网络对图像数据的尺度适应性。但是，这样的改进会引发新的问题。由于卷积核处理的数据来自上一层的输出，使用该方法会增大深度神经网络的宽度，造成卷积计算的复杂度过大，影响特征图的使用效率。为了避免由于宽度增加而带来的网络深度过深和参数过多等问题，将 1×1 的卷积加入到 3×3、5×5 卷积层之前和 3×3 池化层后。改进后的 Inception 模型如图 2.5（b）所示。该结构采用维度不同的池化层获取不同尺度大小的感受野，最后将其拼接，得到不同尺度特征融合的网络结构。该结构不仅改善了网络过拟合的情况，还大大提高了网络的计算性能。

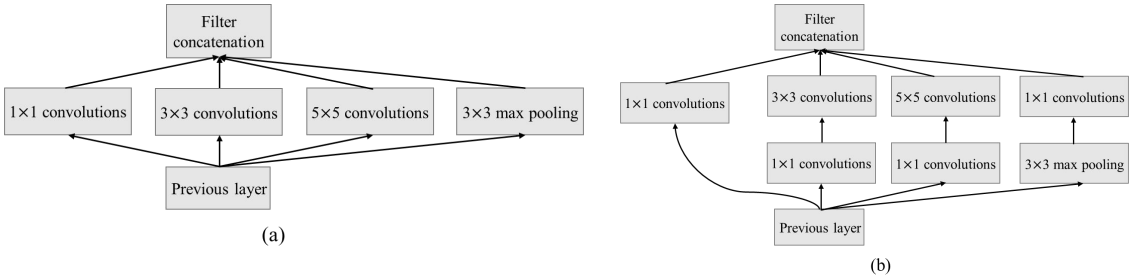

图 2.5　Inception 模型图

2.3 非负矩阵分解算法介绍

非负矩阵分解[49]是指任意给定非负矩阵 $V_{m\times n}$ 和正整数 r（$r \ll \min\{m,n\}$），寻

找非负矩阵$W_{m\times r}$和$H_{r\times n}$，使得：

$$V \approx WH \tag{2-5}$$

其中，r比m或n小，应确保$(m+n)r < mn$，使得W和H的维度小于原矩阵的维度。若$V_{m\times n}$代表预处理后的基因表达数据矩阵，m代表样本数目，n代表基因数目，NMF 算法就是将预处理后的基因数据矩阵分解成特征矩阵$W_{m\times r}$和系数矩阵$H_{r\times n}$，从而实现降维。通常使用目标函数确保 NMF 在分解之后与分解之前的逼近效果，Lee 和 Seung[50]提出了两种目标函数。

基于欧氏距离平方的目标函数：

$$\left\|V-WH\right\|^2 = \sum_{ij}(V_{ij}-(WH)_{ij})^2 \tag{2-6}$$

当且仅当$V=WH$时，式（2-6）得到最优解。

基于广义 KL（Kullback-Leibler）散度的目标函数：

$$D(V\left\|WH\right) = \sum_{ij}(V_{ij}\log\frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij}) \tag{2-7}$$

当且仅当$V=WH$时，式（2-7）得到最小值。

非负矩阵分解是非凸优化问题，并且是 NP 困难问题，很难求出问题的全局最优解。因此，为了求W和H的最优解，各种 NMF 优化算法被提出，如下所述。

2.3.1 基于乘法校正的 NMF 算法

Lee 和 Seung[51]提出了基于乘法校正的 NMF 算法（NMF Based on Multiplicative Update Algorithms），本文将其简称为 NMF_mu。它将梯度下降法和乘性迭代规则结合，有效克服了两个方法的不足之处，并优化了 NMF 算法。算法的具体步骤如下：

（1）初始化矩阵$W \geq 0$和矩阵$H \geq 0$；

（2）分别对矩阵W和矩阵H迭代；

对式（2-6）的更新法则为：

$$W_{ia} \leftarrow W_{ia}\frac{(VH^T)_{ia}}{(WHH^T)_{ia}} \tag{2-8}$$

$$H_{au} \leftarrow H_{au}\frac{(W^TV)_{au}}{(W^TWH)_{au}} \tag{2-9}$$

对式（2-7）的更新法则为：

$$W_{ia} = W_{ia}\frac{\sum_u H_{au}V_{iu}/(WH)_{iu}}{\sum_v H_{av}} \tag{2-10}$$

$$H_{au} = H_{au}\frac{\sum_i W_{ia}V_{iu}/(WH)_{iu}}{\sum_k W_{ka}} \tag{2-11}$$

（3）重复步骤（2），直至收敛。

2.3.2 基于交替最小二乘法的 NMF 算法

非负矩阵分解是非凸优化问题，若将矩阵 W 确定后，对于矩阵 H 来说，非凸优化问题就转变为凸优化问题。简而言之，就是将 W 因子固定之后计算最优因子 H 的凸优化问题。Paatero[52]等人提出了基于交替最小二乘法的 NMF 算法（NMF Based on Alternating Least Square Algorithms），在本文中将其简称为 NMF_als。算法的具体步骤如下：

（1）初始化矩阵 $W \geq 0$；

（2）固定矩阵 W，用式（2-12）更新 W：

$$W \leftarrow \arg \min_{W \in R^{m \times r}, W \geq 0} \left\| V - WH \right\|_F^2 \tag{2-12}$$

（3）固定矩阵 H，用式（2-13）更新 H：

$$H \leftarrow \arg \min_{H \in R^{r \times n}, H \geq 0} \left\| V - WH \right\|_F^2 \tag{2-13}$$

循环直至收敛或达到最大迭代次数。

2.3.3 基于 OBS 和交替最小二乘法的 NMF 算法

Optimal Brain Surgery（OBS）[53,54]算法是一种基于 Hessian 矩阵的网络修剪算法。算法步骤如下：

（1）构造误差曲面的局部模型，分析权值扰动所造成的影响。将误差函数进行 Taylor 展开：

$$\delta E = (\frac{\delta E}{\delta w}) \cdot \delta w + \frac{1}{2} \delta w^T \cdot H \cdot \delta w + O(\left\| \delta w \right\|^3) \tag{2-14}$$

其中，H 为 Hessian 矩阵，T 表示矩阵的转置，w 为神经网络中的参数，E 为训练集的训练误差。该剪枝算法对任意的优化算法都是适用的。

（2）通过拉格朗日乘子法，可求解约束最优化问题。

$$S = \frac{1}{2} \Delta w^T H \Delta w - \lambda (l_i^T \Delta w + w_i) \tag{2-15}$$

其中，λ 是 Lagrange 乘子。利用矩阵的逆，求得权值向量 w 中的最佳变化：

$$\Delta w = -\frac{w_i}{\left[H^{-1} \right]_{i,i}} H^{-1} l_i \tag{2-16}$$

（3）Lagrange 算子 S 对元素 w_i 的相应最优值为：

$$L_i = -\frac{w_i^2}{2\left[H^{-1} \right]_{i,i}} \tag{2-17}$$

其中，H^{-1} 是 Hessian 矩阵的逆，$[H^{-1}]_{i,i}$ 是逆矩阵中第(i,i)个元素。在 OBS 过程中，最小特征值的权值将被删除，剩余权值按照式（2-16）进行校正。基于 OBS 和交

替最小二乘法的 NMF 算法（NMF Based on Optimal Brain Surgery and Alternate Least Square Algorithms），在本文中将其简称为 NMF_alsobs。NMF_alsobs 是基于 OBS 算法来对式（2-12）和式（2-13）中的 W 和 H 进行迭代优化。优化步骤如下：

（1）基于交替最小二乘迭代优化问题，构造误差曲面的局部模型，分析矩阵中负数所造成的影响；

（2）构建 Lagrange 算子解决约束最优化问题。

（3）求得最优的 W 或 H。

2.3.4 基于投影梯度算法的 NMF 算法

因为

$$F(W,H)=\frac{1}{2}\|V-WH\|_F^2=\frac{1}{2}\sum_{i=1}^{n}\|V_{*i}-(WH)_{*i}\|_2^2 \qquad (2\text{-}18)$$

则非负矩阵分解问题可看成在凸集上的 n 个独立非线性优化问题。使用投影梯度法求解如下非线性优化问题：

$$\min_{x\in R_+^n} f(x) \qquad (2\text{-}19)$$

其中，$f(x)$ 是 R^n 上的可微函数。Lin[55]提出了一种交替非负最小平方投影梯度算法，求解式（2-19）。算法的具体步骤如下：

（1）输入：常数 β 和 σ，其中 $0<\beta<1$，$0<\sigma<1$，初始可行点为 x^1；

（2）对于迭代次数 k = 1，2，3，4…，

（a）

$$x^{k+1}=P\left[x^k-\alpha_k\nabla f(x^k)\right] \qquad (2\text{-}20)$$

其中，$\alpha_k=\beta^{t_k}$，t_k 依次取值 1，2，3…，当 α_k 满足式子（2-21）时，t_k 停止取值，将停止时的取值记为 t。

$$f(x^{k+1})-f(x^k)\le\sigma\nabla f(x^k)^T(x^{k+1}-x^k) \qquad (2\text{-}21)$$

（b）检验 x^{k+1} 是否满足式子（2-22）和（2-23）的收敛准则：

$$\left\|\nabla^P f(x^k)\right\|\le\varepsilon\left\|\nabla f(x^1)\right\| \qquad (2\text{-}22)$$

其中，

$$\nabla^P f(x)_i\equiv\begin{cases}\nabla f(x)_i & l_i\le x_i\le u_i\\ \min(0,\nabla f(x)_i) & x_i=l_i\\ \max(0,\nabla f(x)_i) & x_i=u_i\end{cases} \qquad (2\text{-}23)$$

若满足，则输出 $\{x^k\}_{k=1}^{\infty}$，算法停止。若不满足，则重复步骤（2）

2.3.5 概率非负矩阵分解

基因表达数据易受噪声的影响，使得预测的准确性降低。考虑到数据随机性

的特点，对这一类数据进行系统化的分析和处理。Belhassen Bayar 等人[56]提出了概率非负矩阵分解算法（Probabilistic Nonnegative Matrix Factorization，以下简称PNMF），该算法假设数据通过一个多项式概率密度函数得到，即在随机的场景下优化了 NMF 的架构和算法。

PNMF 的目标函数为：

$$R(W,H)=\|V-WH\|_F^2+\alpha\|W\|_F^2+\beta\|H\|_F^2 \tag{2-24}$$

迭代规则如下：

$$W_{ij}\leftarrow W_{ij}\frac{(VH^T)_{ij}}{(WHH^T+\alpha W)_{ij}} \tag{2-25}$$

$$H_{ij}\leftarrow H_{ij}\frac{(W^TV)_{ij}}{(W^TWH+\beta H)_{ij}} \tag{2-26}$$

在式（2-25）和式（2-26）的迭代规则下，式（2-24）的目标函数是非增的，而当 W 和 H 固定在一个点时，函数 R 是固定不变的。

2.4 性能评估

2.4.1 评价方法

使用机器学习处理分类任务时，往往将数据集分成三个部分：训练集、验证集和测试集。训练集用来对机器学习模型进行参数优化从而达到训练目的。验证集通常用来对模型进行评估，调整模型的超参数，对模型进行无偏估计。测试集能够验证和评估机器学习模型的性能。科研工作者使用特定的标准对数据集进行划分，不同的划分方法对最终的分类结果会产生一定程度的影响，常见的数据集划分方法有：留出法和交叉验证法。

（1）留出法

留出法（hold-out）将数据集划分为互不相交的集合。两个集合是互斥的，其中一个作为训练集，另一个作为测试集。训练集用来构建机器学习模型，测试集用来对机器学习模型进行评估。

（2）K 折交叉验证法

K 折交叉验证（K-fold cross validation），亦称为 K 次交叉验证，将初始样本分为数量相等的 K 份，取其中一份样本集作为验证模型的数据，其余 K-1 份样本集用来训练。交叉验证会重复 K 次，并且每个子样本集均被验证过一次，将测试结果求取平均值作为模型性能的评价指标。

2.4.2 评价标准

机器学习模型的性能评估，既要有正确的实验方法，也需要能够量化模型效

果的评价指标。在机器学习中，可以采用混淆矩阵[57]（confusion matrix）作为评价机器学习算法的量化指标，该方法能够较好的反映出模型的性能水平，帮助科研工作者找到适合当前任务的机器学习模型。混淆矩阵中的二分类模型，主要是采用准确率、精确率、召回率、F 值以及 ROC 曲线[58]进行模型评估。

表 2.3　分类结果统计表

	实际属于该类的样本数	实际不属于该类的样本数
分类器判定属于该类的样本数	True Positives(TP)	False Positives(FP)
分类器判定不属于该类的样本数	False Negatives (FN)	True Negatives(TN)

（1）准确率

准确率表示样本中分类正确的数量占样本总数量的比例，记为 *Accuracy*，其计算公式如（2-27）所示。

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2-27}$$

（2）精确率

精确率表示的是模型预测出来的某一类别样本数量与该类样本总量的比值，记为 *Precision*，其计算公式如（2-28）所示。

$$Precision = \frac{TP}{TP + FP} \tag{2-28}$$

（3）召回率

召回率（查全率）表示模型正确分类的正样本占所有正样本的比值，记为 *Recall*，其计算公式如（2-29）所示。

$$Recall = \frac{TP}{TP + FN} \tag{2-29}$$

（4）F 值

精确率指标和召回率指标从不同的层面反映模型分类的效果。然而，在某些情况下这两个评价标准互为矛盾，如某一个指标上升则另一个指标就会下降。通常采用 F 值将这两个评价指标结合起来进行性能综合评价，较好的均衡了两者的矛盾。其计算公式如（2-30）所示。

$$F_\beta = \frac{(\beta^2 + 1) Precision \cdot Recall}{\beta^2 (Precision + Recall)} \tag{2-30}$$

其中，β 是用来调节查准率和查全率权重的参数，一般情况下 β 值取 1，取 1 时的 F_1 值计算公式如（2-31）所示。

$$F_1 = \frac{2 Precision \cdot Recall}{Precision + Recall} \tag{2-31}$$

（5）ROC 曲线

ROC 曲线指的是接收者操作特征。曲线的横坐标为负正类率特异度（FRP），

即模型分类得到的正类中存在的负实例占所有负实例的比值（1-Specificity）；ROC曲线的纵坐标表示真正类率灵敏度（TPR），也就是模型分类预测得到的正类中含有的正实例占所有正实例的比值（Sensitivity）。ROC 曲线下与横轴纵轴共同围成的面积称为 AUC 值，它的取值范围在 0 到 1 之间。

2.5 本章小结

本章详细介绍了深度神经网络的相关知识、乳腺癌组学数据、非负矩阵分解算法和评价深度学习模型性能的评价指标。首先，细致分析了组学数据中的基因表达数据和临床数据。其次，介绍了常用的深度神经网络以及本文使用的多尺度特征融合等深度学习技术。接着描述了五种非负矩阵分解优化算法。最后介绍了本文使用的性能评估方法。

第 3 章 基于非负矩阵分解的 Multi_NMF 特征选择算法

基因表达数据有“维度高，样本少”的特点，且包含大量的无关信息和冗余特征。降低数据维度、提取重要特征成为精准预测的关键。近年来，非负矩阵分解算法由于其特殊的非负性约束，广泛应用于图像识别、文本分类和生物信息学。如，Yuvaraj 等人[59]采用非负矩阵分解方法选择特征基因，并取得较好的分类效果。Jazayeri 等人[60]使用 NMF 对 DNA 甲基化和基因表达数据进行降维，获得了较低的误差。本章以非负矩阵分解算法为研究基础，提出了一种基于非负矩阵分解的 Multi_NMF 方法。实验结果表明，Multi_NMF 方法能选出与乳腺癌生存期密切相关的特征基因，且比单个 NMF 优化算法在特征提取方面更具有优势。

3.1 Multi_NMF 算法

NMF 算法是在不改变原始数据结构的基础上将非负矩阵分解成两个矩阵相乘的形式。由于分解过程不出现负值，且分解结果有较强的可解释性，用 NMF 分析基因表达数据具有应用价值[61]。基于乘法校正的 NMF 算法（NMF_mu）将梯度下降和乘性迭代两种规则巧妙地结合起来，克服了各自的缺点。但在实际应用中，存在局部收敛性得不到保证，性能不稳定等问题。同时，在迭代过程中，还会出现死锁现象。为了保证算法的收敛性，可使用基于交替最小二乘法的 NMF 算法（NMF_als）来优化非负矩阵分解的损失函数，该算法每次迭代都会降低误差，所以结果一定会收敛。基于 OBS 和交替最小二乘法的 NMF 算法（NMF_alsobs）在 NMF_als 的基础上，删除最小特征值的权值，减小误差，从而精确求解 W 和 H。NMF_pg 是求解边界优化问题的经典方法，它的优点是收敛性易于保证，每次迭代利用梯度信息来判断。基于投影梯度的 NMF 算法（NMF_pg）相较于 NMF_mu，有较好的收敛性，能有效避免 NMF_mu 所遇到的死锁现象。但是 NMF_pg 收敛速度慢。概率非负矩阵分解算法（PNMF）规避了基因表达数据在测量或观察过程中产生的误差和噪声。从上述各种 NMF 优化算法的特点可以看出，不同 NMF 优化算法分解所得的特征各有不同。为了更好的获取原基因表达数据的重要特征，本文提出了 Multi_NMF 特征选择算法，融合上述五种 NMF 优化算法，图 3.1 展示了 Multi_NMF 算法的模型示意图。

为了确保数据的质量和可用性，对基因表达数据进行预处理操作。若某个基因对应的数据中，缺失值数量多于患者数量的 10%，则删除该基因及其对应的数据[62]。剩余的缺失值采用近邻填充算法进行填充[63]。接着对基因表达数据进行归一化操作，并对归一化后的数据进行特征选择。本文采用线性函数归一化方法完成基因表达数据的归一化操作。线性函数归一化也称为离差标准化，其本质是对

原始数据进行线性变换，将数据统一映射到[0，1]区间上。转换函数如下：

$$X^{*}=\frac{x-\min}{\max-\min} \tag{3-1}$$

其中，max 为样本数据的最大值，min 为样本数据的最小值。

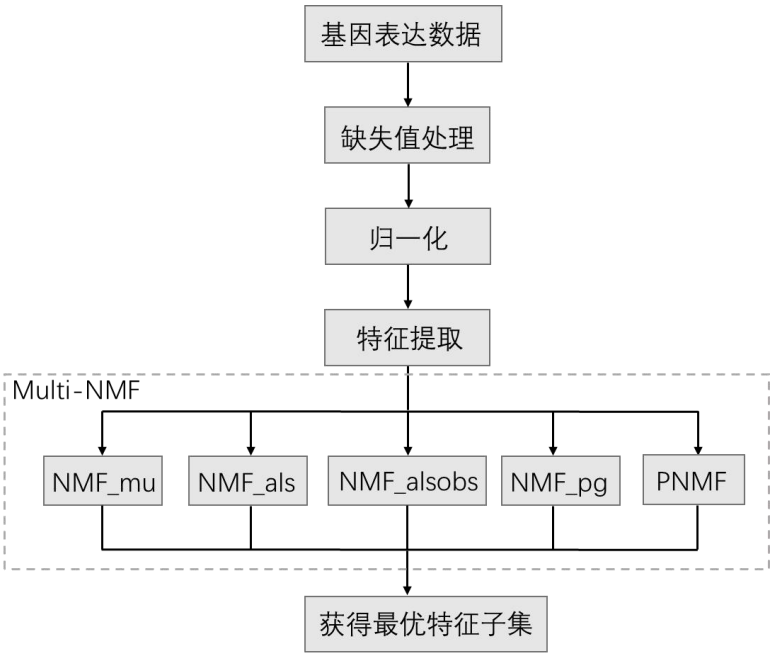

图 3.1　基于非负矩阵分解的 Multi_NMF 模型图

由于单个非负矩阵分解优化算法提取的特征具有局部性，因此本章提出 Multi_NMF 算法来融合五种 NMF 优化算法所提取的特征矩阵。Multi_NMF 方法的主要思想是用 NMF_mu、NMF_als、NMF_alsobs、NMF_pg 和 PNMF 五种 NMF 优化算法对基因表达数据进行特征选择，得到五个特征矩阵。由于单个非负矩阵分解算法所得特征具有片面性，为了获取更全面的基因表达数据特征，本文使用相同的权重将五种 NMF 优化算法所得的特征矩阵进行加权求和，得到融合后的特征子集。最后将这些特征输入神经网络中进行预测。

3.2 实验数据和参数设置

3.2.1 实验数据

本章使用的数据为METABRIC数据集和TCGA乳腺癌数据集中的基因表达数据。作为验证模型基础性能的主要数据集，METABRIC 数据集包含 1980 个样本。依照 Khademi 等人[64]的研究工作，本文也相应地使用 5 年生存期作为划分两类患者的阈值。其中，491 个患者被划分为短生存期样本，1489 个患者被划分为长生存期样本。为了进一步验证 Multi_NMF 方法的有效性和通用性，本文使用 TCGA 数据集来测试模型的性能。TCGA 数据集共有 1196 个样本，其中包含 292 个长期幸存者和 904 个短期幸存者。为了评估算法的性能，本实验将数据集随机划分为 3

组，即 80%的样本做训练集，10%的样本做测试集，剩余 10%的样本做验证集。METABRIC 数据集的划分如表 3.1 所示，TCGA 数据集的划分如表 3.2 所示。其中，训练集用于训练模型，验证集用于调整神经网络模型的参数，测试集用于测试模型性能。本章的实验结果均来自测试集运行 100 次所得结果的平均值。

表 3.1　METABRIC 数据集样本数统计

	长期幸存者	短期幸存者	总计
训练集	1191	393	1584
测试集	149	49	198
验证集	149	49	198
总计	1489	491	1980

表 3.2　TCGA 数据集样本数统计

	长期幸存者	短期幸存者	总计
训练集	234	724	958
测试集	29	90	119
验证集	29	90	119
总计	292	904	1196

3.2.2 参数设置

本文实验基于 Python 编程语言和 TensorFlow 深度学习框架（版本 V1.13.1）。操作系统为 Ubuntu 18.04，GPU 型号为 NVIDIA GeForce GTX 1080，显存 8G，内存 32G，CPU 型号为 Intel(R) Core(TM) i7-7700K CPU@4.20GHz。

在 NMF 算法分解过程中，维度的选取非常关键，它将直接影响所得的特征矩阵，从而导致不同的预测结果。对于 METABRIC 数据集，经 Multi_NMF 方法特征提取后的维度是 200 维。TCGA 数据集经 Multi_NMF 方法降维后的维度为 400 维。

3.3 实验与结果分析

3.3.1 实验设计

为了验证 Multi_NMF 算法的有效性，本章设置对比实验如下：

（1）Multi_NMF 算法与单个 NMF 优化算法进行对比。将预处理后的基因表达数据分别用五种 NMF 优化算法进行特征选择，分解后的特征矩阵输入神经网络进行预测。根据实验结果分析由单个 NMF 特征选择算法所得的预测结果和使用 Multi_NMF 方法进行特征选择所得的预测结果，从而验证 Multi_NMF 算法相较于

单个 NMF 优化算法能提取更加全面的特征信息，具有更好的预测效果。

（2）Multi_NMF 算法与缺少任意一种 NMF 优化算法的组合进行对比。为了验证融合五种 NMF 优化算法的有效性，结合控制变量法的思想来设置对比实验。将预处理后的基因表达数据分别输入任意四种 NMF 优化算法进行特征选择，并将得到的四个特征矩阵赋予相同的权重进行拼接，融合后的特征输入神经网络进行预测。通过此实验可验证，相较于其他几种 NMF 优化算法的组合，Multi_NMF 算法取得了最好的预测结果。

3.3.2 单个 NMF 优化算法性能分析

为了验证本章提出的 Multi_NMF 特征选择算法对乳腺癌生存期预测的有效性，本文将 Multi_NMF 算法分别与五种单个 NMF 优化算法进行比较。五种非负矩阵分解优化算法分别是 NMF_mu、NMF_als、NMF_alsobs、NMF_pg 和 PNMF。本文将这些特征选择算法所得预测结果用 AUC 值、准确率和精确率等评价指标展示。图 3.2 展示了 Multi_NMF 方法与单个 NMF 优化算法在 METABRIC 数据集上的 AUC 值。

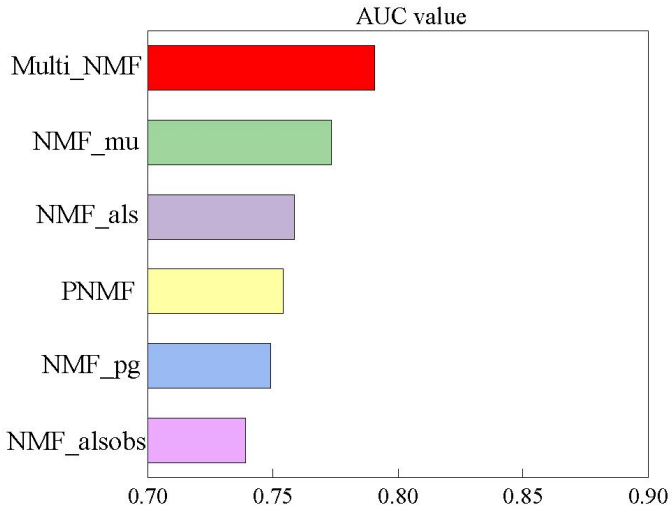

图 3.2　METABRIC 数据集上 Multi_NMF 方法和单个 NMF 优化算法的 AUC 值

由图 3.2 可知，不同 NMF 优化算法所提取的特征不同，有着不同的预测结果。在五种单个 NMF 优化算法中，NMF_mu 取得了较好的预测结果，AUC 值为 77.34%，NMF_alsobs 是这五种 NMF 优化算法中预测性能较差的，AUC 值为 73.93%。作为五种 NMF 优化算法的融合，Multi_NMF 算法的 AUC 值均高于其他 NMF 优化算法，AUC 值为 79.06%。其他五种非负矩阵分解优化算法的 AUC 值分别为 75.86%、73.93%、74.91%、77.34%和 75.44%。相较于单一非负矩阵分解优化算法的预测结果，Multi_NMF 方法分别提高了 3.2%，5.13%，4.15%，1.72%，3.62%。因此，本文提出的 Multi_NMF 方法在特征提取方面具有一定的可行性与有效性，

说明 Multi_NMF 方法有效融合了五种 NMF 优化算法，具有互补的优势。

同时，除了 AUC 值，我们还使用准确率和精确率作为评价指标。实验结果如表 3.3 所示。从实验结果可以看出，Multi_NMF 方法的精准率和准确率均高于其他 NMF 优化算法。其中，Multi_NMF 的 Accuracy 和 Precision 分别为 80.31%和 83.97%，比 NMF_alsobs 的 Accuracy 和 Precision 提高了 3.72%和 3.61%。在五种单个 NMF 优化算法中，NMF_mu 的准确率最高，Accuracy 值为 80%。就精确率而言，PNMF 的 Precision 值高达 83.31%。因此，Multi_NMF 算法的预测效果总体上明显优于其他单个 NMF 优化算法。综上所述，Multi_NMF 方法确实可以弥补单个 NMF 优化算法的信息丢失问题，从而获得更好的特征矩阵，对乳腺癌生存期预测具有重要作用。

表 3.3　METABRIC 数据集上 Multi_NMF 方法与单个 NMF 优化算法性能对比

方法	AUC	Accuracy	Precision
NMF_als	75.86%	78.72%	82.8%
NMF_alsobs	73.93%	76.59%	80.36%
NMF_pg	74.91%	79.78%	83.01%
NMF_mu	77.34%	80.00%	82.82%
PNMF	75.44%	79.78%	83.31%
Multi_NMF	**79.06%**	**80.31%**	**83.97%**

为了更进一步评估 Multi_NMF 方法的泛化能力，本文将 TCGA 乳腺癌数据集作为独立测试集。在 TCGA 数据集上对比 Multi_NMF 算法和单个 NMF 优化算法的性能，实验结果如图 3.3 所示。

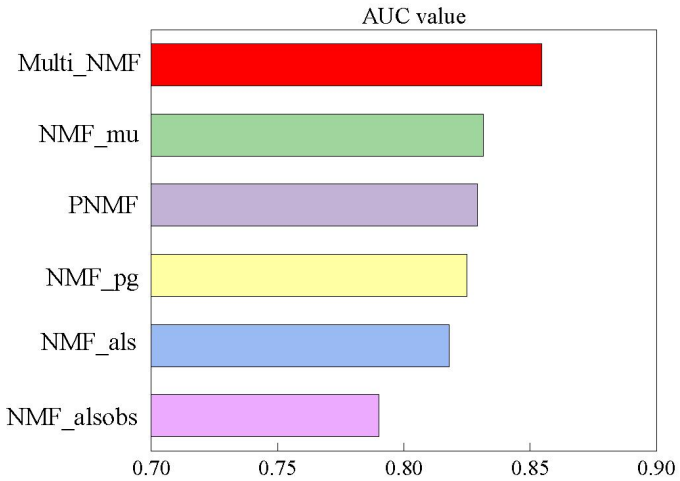

图 3.3　TCGA 数据集上 Multi_NMF 方法和单个 NMF 优化算法的 AUC 值

通过比较实验结果可发现，由于数据集的差异性，导致不同数据集上的实验效果不同。相较于 METABRIC 数据集，Multi_NMF 方法在 TCGA 数据集上效果

更佳。在 TCGA 数据集上，Multi_NMF 算法的 AUC 值为 85.47%，而其他五种 NMF 优化算法的 AUC 值分别为 81.8%、79.02%、82.5%、83.14%和 82.9%。进一步证实了 NMF_NMF 算法比其他 NMF 优化算法的预测效果好。总的来讲，相对于单一非负矩阵分解优化算法的结果，Multi_NMF 方法分别提高了 3.67%，6.45%，2.97%，2.33%，2.57%。由此可看出 Multi_NMF 算法的预测性能确实优于其他 NMF 优化算法。

同理，除了 AUC 值，我们也采用 Accuracy 和 Precision 指标来进一步测试 Multi_NMF 方法的有效性。实验结果如表 3.4 所示。通过比对实验结果我们可以发现，在 TCGA 数据集上，Multi_NMF 方法的准确率和精准率均高于单个 NMF 优化算法。Multi_NMF 方法的准确率为 84.64%，比单个 NMF 优化算法分别提高了 3.26%，3.92%，2.2%，1.39%，2.45%。这与 Multi_NMF 方法在 METABRIC 数据集上得出的结论一致。

表 3.4　TCGA 数据集上 Multi_NMF 方法与单个 NMF 优化算法性能对比

方法	AUC	Accuracy	Precision
NMF_als	81.8%	81.38%	82.71%
NMF_alsobs	79.02%	80.72%	81.23%
NMF_pg	82.5%	82.44%	84.44%
NMF_mu	83.14 %	83.25%	85.13%
PNMF	82.90 %	82.19%	83.44%
Multi_NMF	**85.47%**	**84.64%**	**86.56%**

3.3.3 不同组合的 Multi_NMF 算法性能分析

为了进一步验证使用 Multi_NMF 方法融合五种 NMF 优化算法的有效性，本文采用控制变量法的思想，将缺少任意一种 NMF 优化算法的组合与 Multi_NMF 方法进行比较。表 3.5 展示了五种不同非负矩阵分解优化算法的组合方法与 Multi_NMF 方法的实验结果。此处仍采用 AUC 值、Accuracy 和 Precision 指标评估实验结果。

由表3.5可知，Multi_NMF方法的预测效果明显优于其他组合方法。Multi_NMF方法的AUC值为79.06%。而其他组合方法的AUC分别为74.76%、76.87%、74.34%、76.62%和 76.05%。相较于其他 NMF 优化算法的组合，Multi_NMF 方法分别提高了 4.3%，2.19%，4.72%，2.44%，3.01%。对于精准率和准确率指标，Multi_NMF 方法的效果更好。Multi_NMF 的准确率分别提高了 4.25%，2.12%，2.12%，0.53%，4.26%。综上所述，Multi_NMF 方法的整体性能较好，由此证明 Multi_NMF 特征选择方法包含了更多的特征信息。同时，从实验结果中我们可以看出，融合不一

定就能取得好的效果，也可能会由于权重的分配，而导致融合效果不明显。如，NMF_alsobs+NMF_pg+NMF_mu+PNMF 的 AUC 值为 74.76%，比 NMF_alsobs 的 AUC 值高 0.83%，却比 NMF_pg 的 AUC 值低 0.15%。由此可见，融合算法的效果与各 NMF 优化算法的权重有重大关系，这一问题我们将在下一章继续讨论。

表 3.5　METABRIC 数据集上非负矩阵分解组合算法性能对比

方法	AUC	Accuracy	Precision
NMF_alsobs+NMF_pg+NMF_mu+PNMF	74.76%	76.06%	82.23%
NMF_als+NMF_pg+NMF_mu+PNMF	76.87%	78.19%	83.55%
NMF_als+NMF_alsobs+NMF_mu+PNMF	74.34%	78.19%	78.19%
NMF_als+NMF_alsobs+NMF_pg+PNMF	76.62%	79.78%	79.66%
NMF_als+NMF_alsobs+NMF_pg+NMF_mu	76.05%	76.05%	83.22%
Multi_NMF	**79.06%**	**80.31%**	**83.97%**

为了验证 Multi_NMF 方法的通用性，同样在 TCGA 数据集上测试 Multi_NMF 方法与其他 NMF 组合算法的性能。实验结果如表 3.6 所示。

表 3.6　TCGA 数据集上非负矩阵分解组合算法性能对比

方法	AUC	Accuracy	Precision
NMF_alsobs+NMF_pg+NMF_mu+PNMF	83.36%	81.53%	84.61%
NMF_als+NMF_pg+NMF_mu+PNMF	84.13%	82.27 %	85.03%
NMF_als+NMF_alsobs+NMF_mu+PNMF	83.41%	82.57 %	83.33%
NMF_als+NMF_alsobs+NMF_pg+PNMF	82.27%	80.83 %	82.97%
NMF_als+NMF_alsobs+NMF_pg+NMF_mu	83.51%	81.32 %	83.75%
Multi_NMF	**85.47%**	**84.64%**	**86.56%**

由表 3.6 可知，Multi_NMF 方法的整体性能明显高于其他 NMF 优化算法的组合。其中，Multi_NMF 方法的 AUC 值为 85.47%，而其他组合方法的 AUC 分别为 83.36%、84.13%、83.41%、82.27%和 83.51%。相较于其他非负矩阵分解优化算法组合的结果，Multi_NMF 方法分别提高了 2.11%、1.34%、2.06%、3.20%和 1.96%。就精准率和准确率指标来说，Multi_NMF 方法的效果更好。Multi_NMF 的准确率分别提高了 3.11%、2.37%、2.07%、3.81%和 3.32%。由此证明，Multi_NMF 方法相较于其他NMF优化算法的组合能提取更优质的基因特征，具有更好的预测效果。

3.4 本章小结

本章在 NMF 算法的理论基础上，分析了 NMF_mu、NMF_als、NMF_alsobs、NMF_pg 和 PNMF 五种 NMF 优化算法。鉴于单一非负矩阵分解算法的分解结果具有局部性，未能获取全面的特征，本章在此基础上提出了一种融合五种 NMF 优化

算法的 Multi_NMF 方法。实验结果表明，Multi_NMF 方法具有单个 NMF 优化算法的互补优势，从而取得较好的预测效果。与同类组合方法相比，Multi_NMF 方法比其他 NMF 组合方法有更好的预测性能。综上所述，Multi_NMF 方法有较好的特征选择能力。

第 4 章 基于组学数据和注意力机制的生存期预测模型

上一章提出了基于非负矩阵分解算法改进的 Multi_NMF 方法，实验结果表明，该方法具有较好的特征选择能力。然而上述研究依然存在局限性，如：实验数据仅使用单一的基因表达数据，并未使用包含生物特征的临床数据；模型使用相同权重融合五种 NMF 算法，并未区分不同 NMF 优化算法的差异性等。研究表明[65]，临床数据中含有与乳腺癌生存期相关的重要特征，因此融合组学数据为乳腺癌生存期预测研究提供了新思路。深度学习是近年来机器学习领域的重大突破，大量研究人员将深度学习技术用于乳腺癌生存期预测[66]，并取得较好的预测结果。受上述研究启发，本章探索基于深度学习的组学数据融合方法，并提出一种基于组学数据和注意力机制的深度神经网络模型用于乳腺癌生存期预测。该模型不仅考虑了组学数据，还利用注意力机制自适应地融合了多种特征提取方法，进一步提升了乳腺癌生存期预测准确率。

4.1 融合组学数据的 Attention 机制深度神经网络模型

本章提出了一种融合组学数据的 Attention 机制深度神经网络模型（Attention-based Multi-Nmf Deep neural network using omic data，AMND）来预测乳腺癌患者的生存期。第三章中的 Multi_NMF 方法将五种 NMF 优化算法提取的特征进行融合，取得了较好的预测结果。然而，直接加权求和会导致每个特征矩阵的权重相同，这种做法不是最有效的。因此，本章引入 Attention 机制，提出 AMND 方法来解决这个问题。该模型自适应地融合了五种 NMF 优化算法，得到新的基因特征，并将其与临床数据特征进行融合，放入神经网络中预测。AMND 的结构如图 4.1 所示。

首先，本文使用 NMF_mu、NMF_als、NMF_alsobs、NMF_pg 和 PNMF 算法对基因表达数据进行特征提取，得到五个特征矩阵。然后，采用 Attention 机制对每种 NMF 优化算法加权求和。与 Multi_NMF 方法不同的是，Attention 机制根据患者的临床数据特征，计算出每种 NMF 优化算法的权重。将上述五种 NMF 优化算法中第 j 个 NMF 算法得到的特征矩阵，其中的任一样本 x_i 分别记为 $F_j\_x_i$，该样本的临床特征向量记为 C_x_i。权重计算公式如（4-1）所示。

$$w_j^i = \left(F_j\_x_i\right)^T W \left(C\_x_i\right) \tag{4-1}$$

则，$F_1\_x_i$、$F_2\_x_i$、$F_3\_x_i$、$F_4\_x_i$ 和 $F_5\_x_i$ 分别和 C_x_i 输入神经网络可以得到 w_1^i、w_2^i、w_3^i、w_4^i 和 w_5^i。得到的权重用 *softmax* 函数进行归一化，例如，归一化后的 $\widehat{w_j^i}$ 可以看做第 j 个 NMF 优化算法对第 i 个样本的贡献大小。$\widehat{w_j^i}$ 的计算公式如（4-2）所示。

$$\widehat{w_j^i} = \frac{e^{w_j^i}}{\sum_1^n e^{w_j^i}} \tag{4-2}$$

最后，对五种 NMF 优化算法所得的特征矩阵进行加权求和得到 F，F 的计算公式如（4-3）所示。

$$F = \widehat{w_1^i} F_1\_x_i + \widehat{w_2^i} F_2\_x_i + \widehat{w_3^i} F_3\_x_i + \widehat{w_4^i} F_4\_x_i + \widehat{w_5^i} F_5\_x_i \tag{4-3}$$

然而式（4-3）得到的 F 实际上仅包含了基因表达谱数据，为了考虑组学数据，将 F 与临床数据融合，融合后的组学数据放入神经网络进行预测。AMND 是一个端到端的模型，其中，DNN 的参数可通过训练来进行优化和调整。

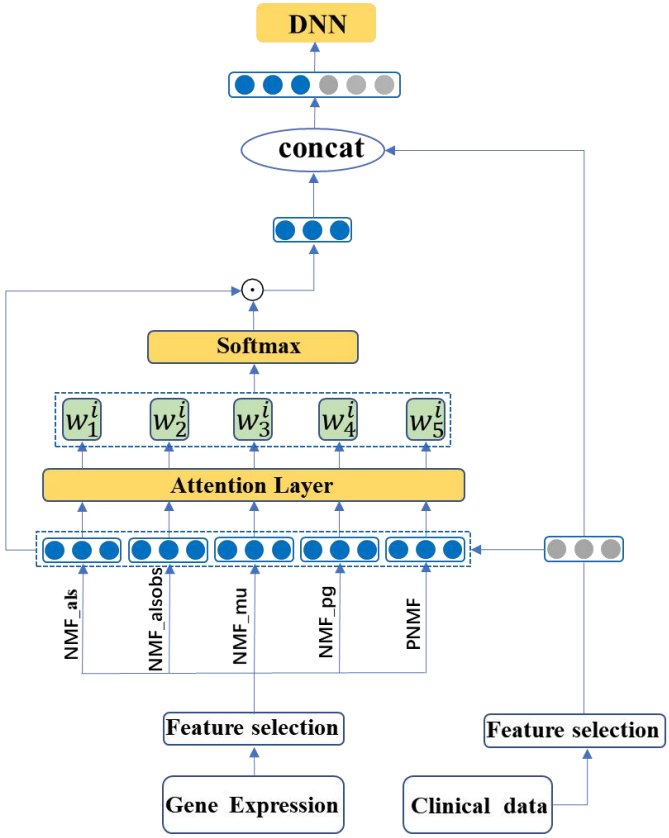

图 4.1　融合组学数据的深度神经网络模型图

4.2 实验数据和参数设置

4.2.1 实验数据

本章使用的组学数据为METABRIC 数据集和 TCGA 数据集中的基因表达数据和临床数据。基因表达数据的预处理流程与第三章一致。METABRIC 数据集中，每个病例有 27 维临床特征，如组织学分级、肿瘤大小、细胞转移状态以及病理分期等。去掉缺失值多的特征数据，本文最后保留 25 维临床信息。与第三章一样，为了保证划分的数据集中，正例样本和负例样本的分布一致，本文将短生存期样本和长生存期样本分别以 8:1:1 的比例随机划分，再将其整合，形成训练集，测试

集和验证集。

4.2.2 参数设置

优化超参数是训练神经网络模型的关键，选择恰当的超参数能够显著改善模型的学习能力，进而大幅度提升模型性能。在本文中，AMND 模型的超参数主要包含初始学习率、隐藏层数量、最小批量尺寸、激活函数等。表 4.1 展示了 AMND 模型的超参数设置。

表 4.1　AMND 模型的超参数设置

参数	参数值
隐藏层数量	3
隐藏层神经元数量	200
学习率	10^{-3}
最小批量尺寸	32
激活函数	Tanh
优化器	Adam

4.3 实验与结果分析

4.3.1 实验设计

为了验证本章提出的 AMND 模型在乳腺癌生存期预测方面的性能，本章在 METABRIC 数据集和 TCGA 数据集上，设置了多角度的对比实验，如组学数据与单一数据的对比、Attention 机制效果的对比等。对比实验设计如下：

（1）AMND 模型与基于单个 NMF 优化算法的深度神经网络模型对比。在 AMND 模型中，通过 Attention 机制计算出的权重来融合五种 NMF 优化算法所得的特征矩阵。而基于单个 NMF 优化算法的深度神经网络模型中所使用的特征来自于单个 NMF 优化算法。本实验可验证 Attention 自适应融合五种 NMF 优化算法的有效性。

（2）AMND 模型与基于不同数据的深度神经网络模型对比。为了验证组学数据对乳腺癌生存期预测的有效性。本章设置了仅使用基因表达数据或者仅使用临床数据作为模型的输入，来进行乳腺癌生存期预测。本实验不仅能证实组学数据相较于单一组学数据的有效性，还能进一步分析基因表达数据和临床数据对乳腺癌生存期预测的影响。

（3）临床数据的作用点分析。由图 4.1 可知，临床数据在 AMND 模型中使用了两次，第一次是通过 Attention 机制计算权重，第二次是融合组学数据。本章修改 AMND 模型，分别设置临床数据仅用于 Attention 机制计算权重和临床数据仅用

于融合组学数据的实验。本实验用于验证临床数据在 AMND 模型中所起的重要作用。

（4）AMND 模型与传统的机器学习方法进行比较。在分类任务中，分类器的选择至关重要。本章设计了 AMND 方法与 SVM、LR 和 RF 三种方法的对比实验。在 AMND 模型中，使用全连接神经网络进行预测，此处将预测方法分别设置为 SVM、LR 和 RF 等经典的机器学习方法。本实验可对比深度神经网络和传统机器学习方法在乳腺癌生存期预测中的效果。

4.3.2 基于单个 NMF 优化算法的深度神经网络模型性能分析

为了验证 Attention 机制自适应融合 Multi_NMF 算法的有效性，本章采用基于单个 NMF 优化算法的深度神经网络模型进行预测。AMND 模型使用 NMF_mu、NMF_als、NMF_alsobs、NMF_pg 和 PNMF 算法所得的特征矩阵与临床数据计算相似度，并用 Attention 机制计算的相似度权重将五种 NMF 优化算法加权求和，融合后的基因表达数据与临床数据进行二次拼接，拼接后的组学数据输入神经网络中进行预测。而基于单个 NMF 优化算法的深度神经网络模型是直接将单个 NMF 优化算法（如，alsobs 算法）所提取的特征与临床数据进行融合，最后将融合后的组学数据特征向量输入神经网络中进行预测。

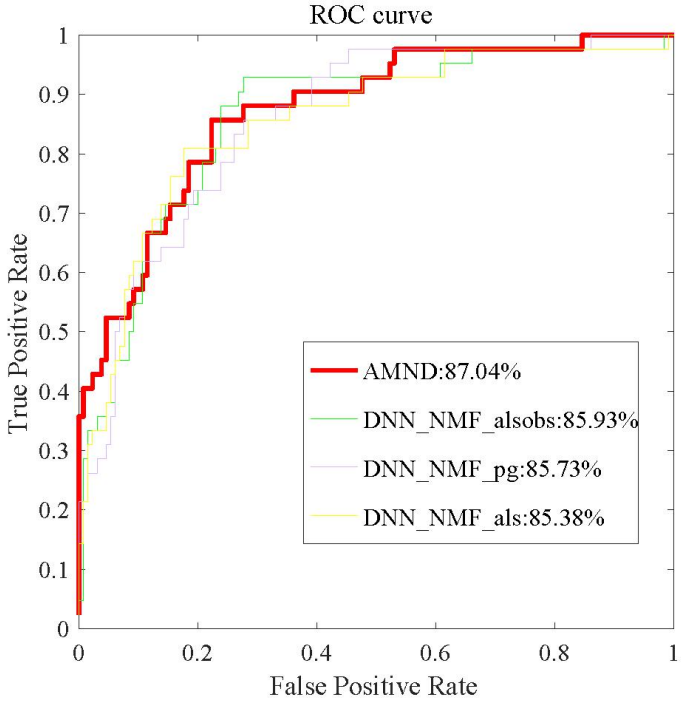

图 4.2　METABRIC 数据集上 AMND 和基于单个 NMF 优化算法的模型的 ROC 曲线

此处将 NMF_mu、NMF_als、NMF_alsobs、NMF_pg 和 PNMF 算法结合深度神经网络的模型分别命名为 DNN-NMF_mu、DNN-NMF_als、DNN-NMF_alsobs 、DNN-NMF_pg 和 DNN_PNMF。图 4.2 展示了 AMND 与 DNN-NMF_alsobs、

DNN-NMF_pg 和 DNN-NMF_als 的 ROC 曲线。由图 4.2 可知，相较于单个 NMF 优化算法的深度神经网络模型，AMND 方法取得了最好的 AUC 值（87.04%），具有更好的预测效果。除了 ROC 曲线外，每种方法相应的 AUC 值也被计算出来。如图 4.3 所示，DNN-NMF_alsobs、DNN-NMF_pg、DNN-NMF_als、DNN_PNMF 和 DNN-NMF_mu 的 AUC 值分别为 85.93%、85.73%、85.38%、84.95%和 85.35%。AMND 比 DNN_PNMF 、 DNN-NMF_mu 、 DNN-NMF_als 、 DNN-NMF_pg 和 DNN-NMF_alsobs 的 AUC 值分别提升了 2.09%、1.69%、1.66%、1.31%和 1.11%。由此证实了基于 Attention 机制的神经网络能自适应地调整各 NMF 优化算法的权重，且融合组学数据能有效提升乳腺癌生存期预测性能。

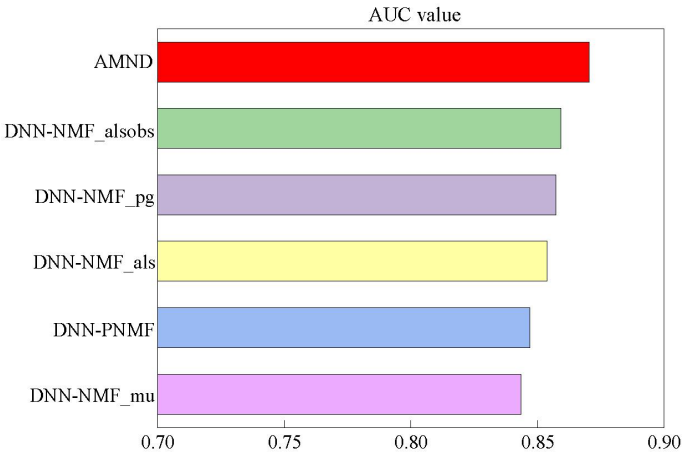

图 4.3　METABRIC 数据集上 AMND 和基于单个 NMF 优化算法的模型的 AUC 值

除了 AUC 值，我们还使用 Acc、Pre、F1-score 和 Recall 等性能指标评估模型效果。实验结果如图 4.4 所示。从图 4.4 中，我们可以看出基于不同 NMF 优化算法的深度神经网络模型具有不同的预测效果。在五种单个 NMF 优化算法的模型中，DNN-NMF_pg 取得较好的预测准确率和精确率，DNN-NMF_als 有较高的 F 值和召回率。相较于单个 NMF 优化算法的深度神经网络模型，AMND 方法的准确率、精确率、F1-Score、和召回率明显高于单个 NMF 优化算法的深度神经网络模型。由此可见，组学数据中包含更多的生物特征，能大幅度提升乳腺癌生存期预测准确率。

同时，六种方法对应的 Acc、Pre、F-score 和 recall 数值如表 4.2 所示。AMND 获得了最高的精确率：85.76%。而 DNN-NMF_alsobs、DNN-PNMF 和 DNN-NMF_pg 的 Pre 值为 82.11%、82.99%和 84.72%。AMND 比 DNN-NMF_alsobs，DNN_PNMF，DNN-NMF_pg 分别提升了 3.65%，2.77%，1.04%。结果说明 AMND 对正负例样本的预测准确率比其他五种模型高。就准确率而言，AMND 方法也提升了不少，AMND 比 DNN-NMF_alsobs，DNN-NMF_pg，DNN-NMF_als 分别提升了 4.07%、2.33%和 3.49%。以上所有结果表明，AMND 方法比基于单个 NMF 优化算法的深

度神经网络模型的整体性能更好。证实了基于 Attention 机制的深度神经网络，在乳腺癌生存期预测中的作用显著。

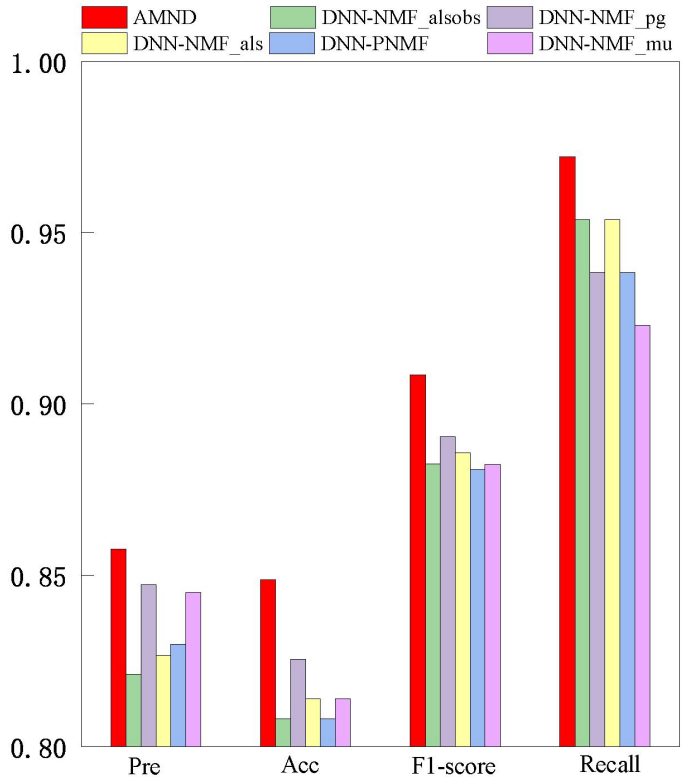

图 4.4 METABRIC 数据集上 AMND 和单个 NMF 优化算法的模型的整体性能对比

表 4.2　METABRIC 数据集上 AMND 和单个 NMF 优化算法模型的整体性能对比

Method	Acc	Pre	F1-score	recall
DNN-NMF_mu	81.39%	84.5%	88.23%	92.3%
DNN-NMF_alsobs	80.81%	82.11%	88.25%	95.38%
DNN-NMF_pg	82.55%	84.72%	89.05%	93.84%
DNN-NMF_als	81.39%	82.66%	88.57%	95.38%
DNN-PNMF	80.81%	82.99%	88.08%	93.84%
AMND	**84.88%**	**85.76%**	**90.84%**	**97.23%**

为了进一步验证 AMND 方法的性能，我们在 TCGA 数据集上进行 AMND 和基于单个 NMF 优化算法的深度神经网络模型的性能对比，图 4.5 展示了 TCGA 数据集上 AMND 和基于单个 NMF 优化算法的模型的 ROC 曲线。从图 4.5 中我们可以看出 AMND 方法的预测效果明显比 DNN-NMF_alsobs、DNN-NMF_pg 和 DNN-NMF_als 方法好。

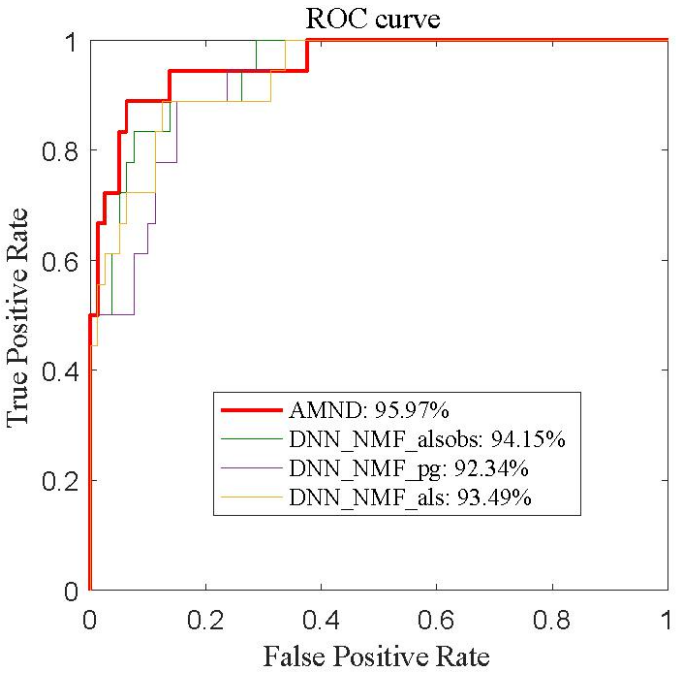

图 4.5　TCGA 数据集上 AMND 和基于单个 NMF 优化算法的模型的 ROC 曲线

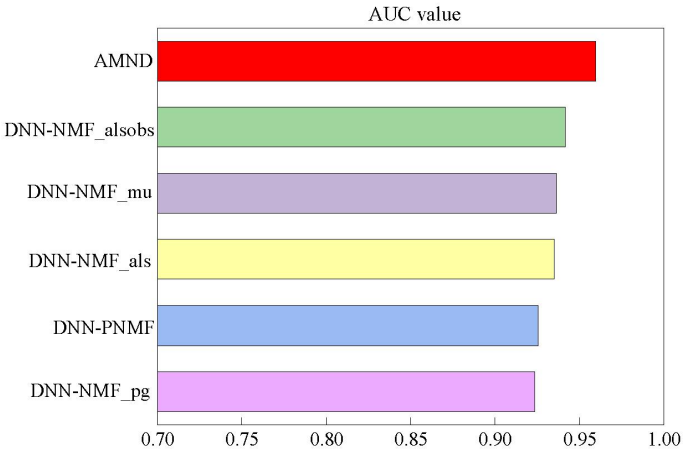

图 4.6　TCGA 数据集上 AMND 和基于单个 NMF 优化算法的模型的 AUC 值

除了 ROC 曲线，图 4.6 描述了 AMND 算法和基于单个 NMF 优化算法的模型的 AUC 值。从图 4.6 可以看出 AMND 算法的预测性能取得了较高的提升，AUC 值高达 95.97%，比 DNN-NMF_alsobs、DNN-NMF_pg 和 DNN-NMF_als 方法提高了 1.82%、3.63%和 2.84%。总体上的结果与 METABRIC 数据集上的结果保持一致。由此可见，Atenntion 机制能够自适应地融合多种 NMF 优化算法，比单个 NMF 优化算法的深度神经网络模型和 Multi_NMF 方法具有更好的预测效果。

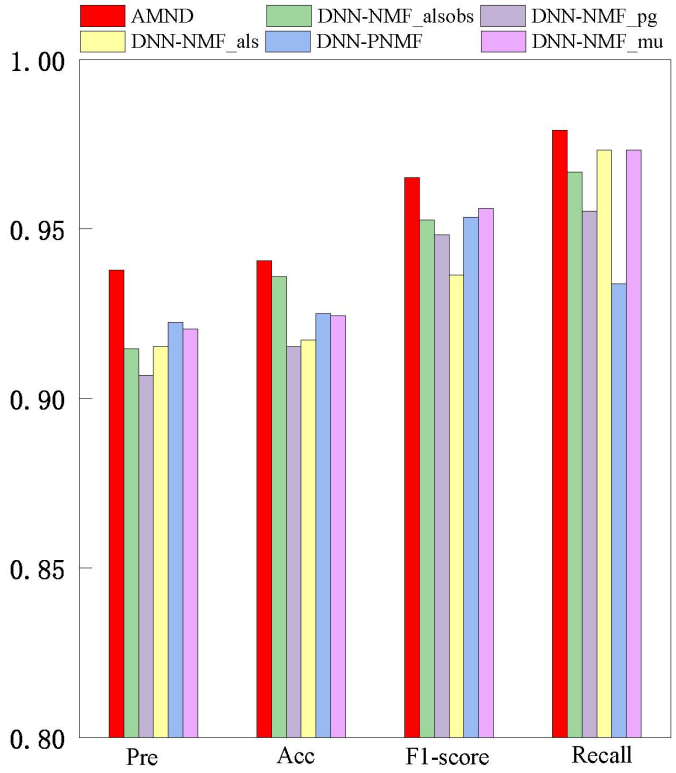

图 4.7　TCGA 数据集上 AMND 和单个 NMF 优化算法的模型的整体性能指标值

同理，除了 AUC 值，我们还将使用 Acc、Pre、F1-score 和 recall 等性能指标进实验对比。实验结果如图 4.7 所示。从图 4.7 中，我们可以看出 AMND 的准确率、精确率、F1-Score、和召回率均高于单个 NMF 优化算法的深度神经网络模型。在五种单个 NMF 优化算法的模型中，DNN-NMF_alsobs 方法整体上比其他 NMF 优化算法的模型预测效果好。

表 4.3　TCGA 数据集上 AMND 和单个 NMF 优化算法的模型的整体性能指标值

Method	Acc	Pre	F1-score	recall
DNN-NMF_mu	92.06%	92.44%	95.61%	97.32%
DNN-NMF_alsobs	91.47%	93.6%	95.27%	96.69%
DNN-NMF_pg	90.69%	91.53%	94.82%	95.53%
DNN-NMF_als	91.53%	91.73%	93.64%	97.32%
DNN-PNMF	92.24%	92.5%	95.35%	93.38%
AMND	**93.79%**	**94.06%**	**96.52%**	**97.91%**

六种方法对应的 Acc、Pre、F-score 和 recall 数值如表 4.3 所示。AMND 获得了最高的精确率：94.06%。而 DNN-NMF_als、DNN_PNMF 和 DNN-NMF_pg 的 Pre 值为 91.73%、92.5%和 91.53%。AMND 比 DNN-NMF_als，DNN-PNMF，DNN-NMF_pg 分别提升了 2.33%，1.56%，2.53%。从实验结果可以看出，AMND 对正反例样本的预测准确率比其他五种模型高。就准确率而言，AMND 方法也提

升了不少，AMND 比 DNN-NMF_alsobs，DNN-NMF_pg，DNN-NMF_als 分别提升了 2.32%、3.10%和 2.26%。以上所有的比较结果表明，AMND 比基于单个 NMF 优化算法模型的整体性能更好。证实了 Attention 机制能够自适应地融合多种特征提取算法，在乳腺癌生存期预测中作用显著。

4.3.3 基于不同数据的深度神经网络模型性能分析

为了证实组学数据有助于提升乳腺癌生存期预测的准确率，本小节仅使用基因表达数据或者仅使用临床数据作为模型的输入，来预测乳腺癌患者的生存期。基于临床数据的深度神经网络模型是直接将临床数据提取后的特征输入神经网络中进行预测，本文将此实验命名为 Only_clinical。基于基因表达数据的深度神经网络模型将基因表达数据用 Multi_NMF 进行特征提取，提取后的特征输入神经网络中进行预测，本文将此实验命名为 Only_exp。本实验不仅能验证组学数据相较于单一数据的有效性，还能进一步分析基因表达数据和临床数据对乳腺癌生存期预测的影响。

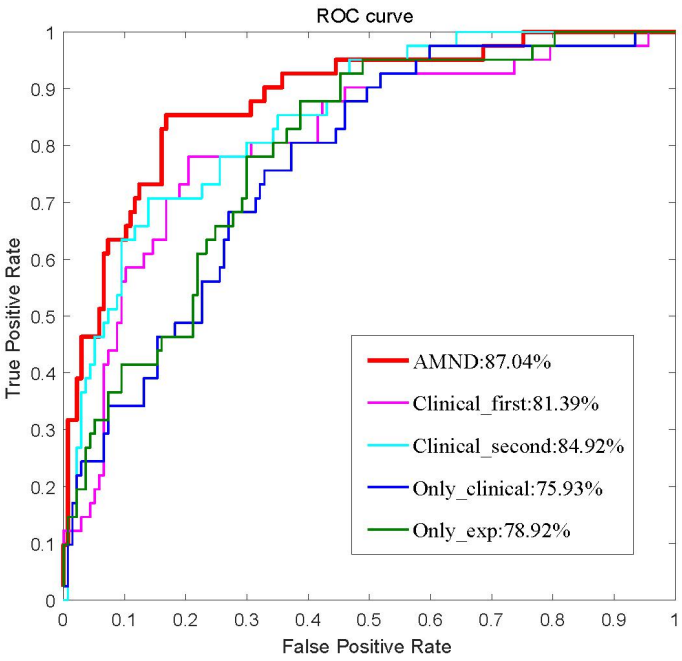

图 4.8　METABRIC 数据集上 AMND 和其他变体模型的 ROC 曲线

图 4.8 展示了 METABRIC 数据集上 AMND 和 Only_clinical、Only_exp、Clinical_first、Clinical_second 模型的 ROC 曲线。从实验结果中我们可以看出，AMND 方法的 ROC 曲下面积明显优于基于单一数据的深度神经网络模型，说明组学数据中包含了更多生物学信息，能有效提高乳腺癌生存期预测准确率。相较于 Only_clinical 和 Only_exp 方法，AMND 方法的预测性能提升了 11.11%和 8.12%。另外，我们也可以发现，临床数据确实对预后有直接影响，但是效果不明显。针

对基因表达数据和临床数据而言，基于基因表达数据的模型效果比临床数据更优，也再次说明了提取基因表达数据特征的重要性和 Multi_NMF 特征提取方法的有效性。综上所述，基于深度学习的组学数据融合方法，包含了更多的生物学特征信息，能有效提升乳腺癌生存期预测准确率。

4.3.4 临床数据作用点性能分析

由 AMND 的模型图（见图 4.1）可知，临床数据在 AMND 模型中使用了两次。第一次是用于 Attention 机制计算权重，第二次用于融合组学数据。为了进一步验证临床数据在这两次中的作用，我们分别设置了第一次使用临床数据（本文将其简称为 Clinical_first）和第二次使用临床数据（本文将其简称为 Clinical_second）的对比实验。实验结果如图 4.8 所示。从图 4.8 中，我们可以得出以下结论：AMND 取得了最好的效果，AUC 值达到了 87.04%，相较于 Clinical_second、Clinical_first、Only_clinical 和 Only_exp 方法，分别提高了 2.12%、5.95%、11.11%和 8.12%。通过 Clinical_first 和 Clinical_second 的结果可知，AMND 的良好效果与两次临床数据的使用密不可分。也就是说，Attention 机制和融合组学数据都能改善乳腺癌生存期预测效果。通过 Clinical_first 和 Only_exp 可知，用 Attention 机制来对五种 NMF 算法加权求和所得到的特征向量确实比基于同等权重加权求和得到的特征向量更具有代表性，从而证实了 Attention 机制这种自适应计算权重的方法能更好的融合五种 NMF 优化算法得到的特征向量，从而得到更好的特征表示。

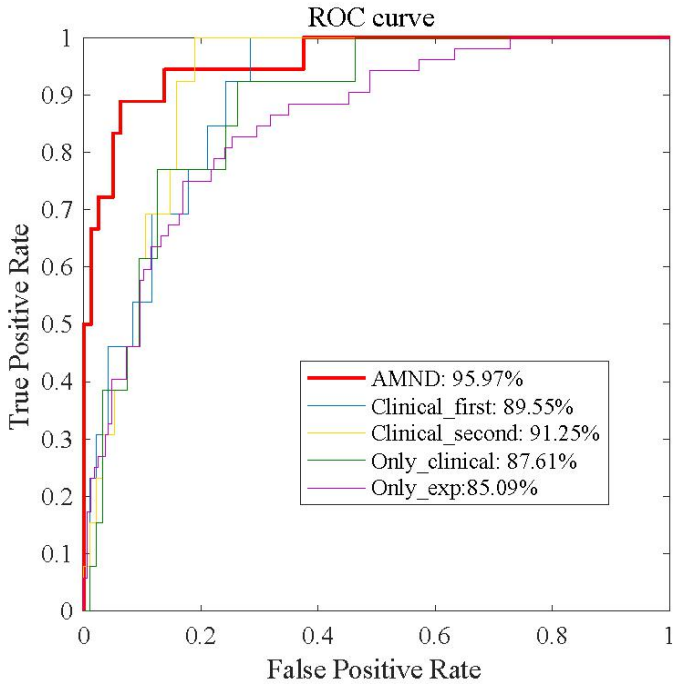

图 4.9　TCGA 数据集上 AMND 和其他变体模型的 ROC 曲线

同样，我们采用 TCGA 数据集作为独立测试集，进一步验证 AMND 模型的有

效性。图 4.9 展示了 TCGA 数据集上 AMND 和其他变体模型的 ROC 曲线。从图 4.9 中可以看出，AMND 方法的预测效果明显优于其他方法，AUC 值达到了 95.97%，相较于 Clinical_second、Clinical_first、Only_clinical 和 Only_exp 方法，分别提高了 4.72%、6.42%、8.36%和 10.88%。从 Only_clinical 和 Only_exp 可以看出， Only_clinical 方法比 Only_exp 方法的 ROC 值高 2.52%，说明临床数据的作用比基因表达数据更明显。Clinical_first 和 Clinical_second 的结果表明，AMND 的良好效果与两次临床数据的使用密不可分。也就是说，Attention 机制和融合组学数据都能改善乳腺癌生存期预测效果。通过 Clinical_first 和 Only_exp 可知，用 Attention 机制来对五种 NMF 算法加权求和所得到的特征向量确实比基于同等权重加权求和得到的特征向量更具有代表性，进而说明了 Attention 机制自适应融合特征选择方法的有效性。

4.3.5 现有方法性能比较

为了进一步验证基于深度学习的生存期预测模型的有效性，本文将 AMND 方法与 SVM、LR 和 RF 三种机器学习方法进行对比。图 4.10 展示了四种方法的 ROC 曲线，从图 4.10 中可以看出，AMND 方法的曲线下面积均高于其他几种方法。

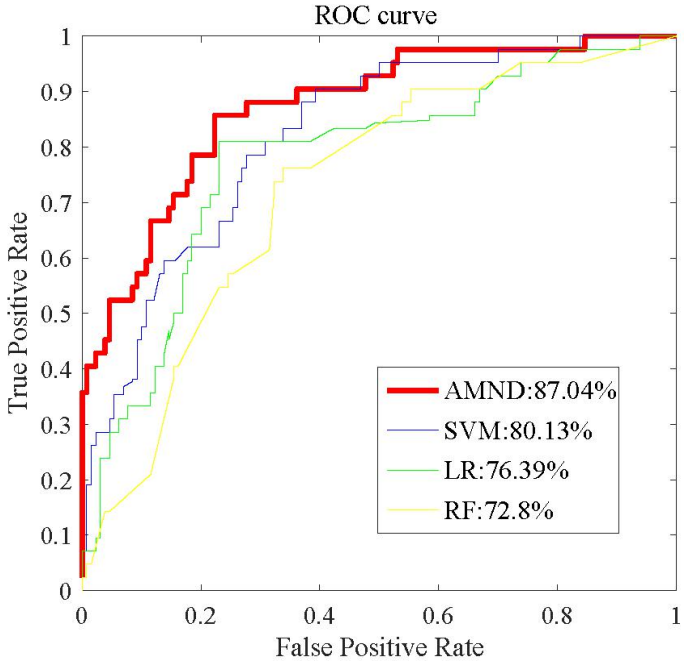

图 4.10　METABRIC 数据集上 AMND 算法与 SVM、LR 和 RF 方法的 ROC 曲线

除了 ROC 曲线外，每种方法对应的 AUC 值也被计算出来，如图 4.11 所示。SVM、LR、和 RF 三种方法的 AUC 值分别为 81.9%、77.41%和 73.6%。而 AMND 的方法的 AUC 值为 87.45%，比其他三种方法分别提升了 5.55%，10.04%，13.85%。说明深度学习方法更有利于乳腺癌生存期预测研究。在这三种传统的分类算法中，支持向量机的预测效果最好。通过分析实验结果，我们可知融合组学数据能极大

的提升乳腺癌生存期预测性能，同时 AMND 方法能够更好的利用多种特征提取方法来提升生存期预测准确度。

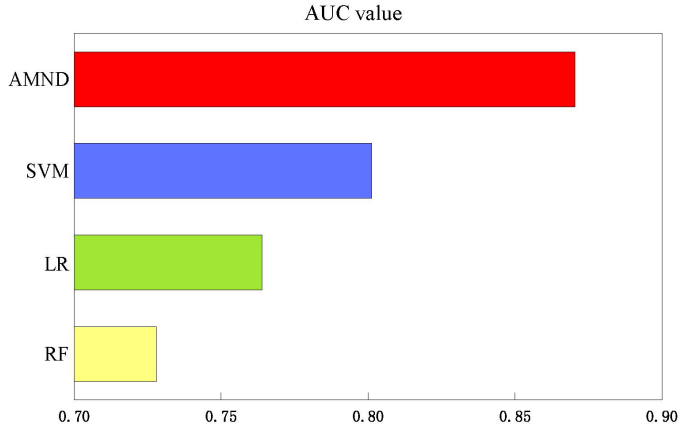

图 4.11　METABRIC 数据集上 AMND 算法与 SVM、LR 和 RF 方法的 AUC 值

除了 AUC 指标，本文还分析了不同方法的准确率、精准率、F1 值和召回率等指标的值，实验结果如图 4.12 和表 4.4 所示。

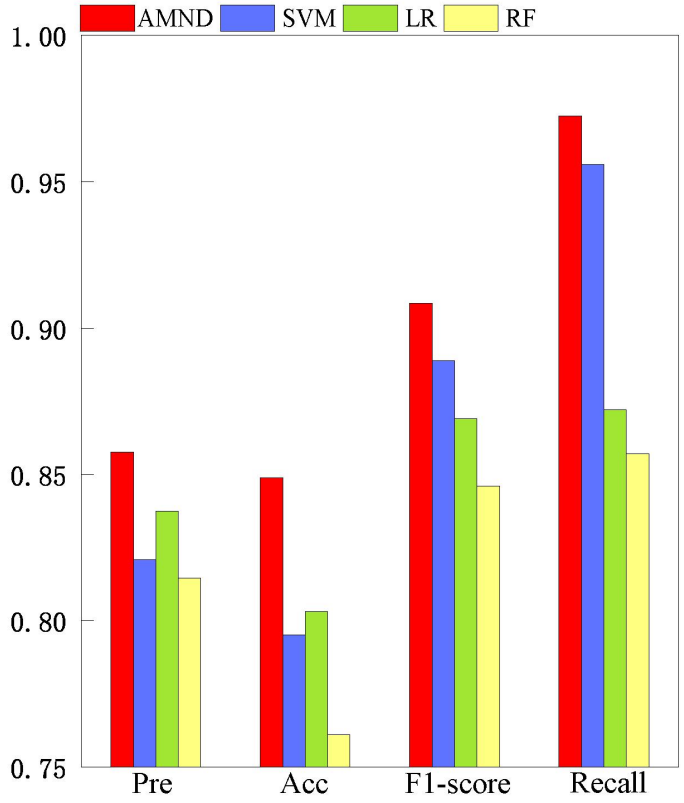

图 4.12　METABRIC 数据集上 AMND 算法与 SVM、LR 和 RF 方法的预测性能

由图 4.12 可知，AMND 方法在 Acc、Pre、F1-score 和 recall 上的性能均高出其他三种方法。从表 4.4 可以看出，AMND 在 Acc、Pre、F1-score 和 recall 上分别高出 SVM 方法 5.23%、2.85%、3.3%和 3.85%。同时，相较于 RF 和 LR 两种方法，

AMND 方法的整体性能得到了改善。总而言之，通过多个性能指标来评估 AMND 方法的性能，可看出 AMND 方法均优于其他几种方法，进一步证实了基于深度学习的方法在乳腺癌生存期预测上的有效性。

表 4.4　METABRIC 数据集上 AMND 算法与 SVM、LR 和 RF 方法的预测性能

Method	Acc	Pre	F1-score	recall
SVM	79.51%	82.08%	88.88%	95.60%
LR	80.31%	83.75%	86.90%	87.20%
RF	76.1%	81.45%	84.60%	85.70%
AMND	**84.88%**	**85.76%**	**90.84%**	**97.23%**

同理，我们在 TCGA 数据集上进一步对比 AMND 方法和传统机器学习方法的优劣。图 4.13 展示了 TCGA 数据集上 AMND 算法与 SVM、LR 和 RF 方法的 ROC 曲线。从图 4.13 可知，TCGA 数据集上的结果与 METABRIC 数据集上的结果一致。AMND 方法的预测效果比 SVM、RF 和 LR 方法好。相较于 SVM、RF 和 LR 方法，AMND 的 AUC 值提高了 9.54%、14.56%和 12.74%。因此，可以看出深度学习的方法比传统机器学习方法的预测效果更好。

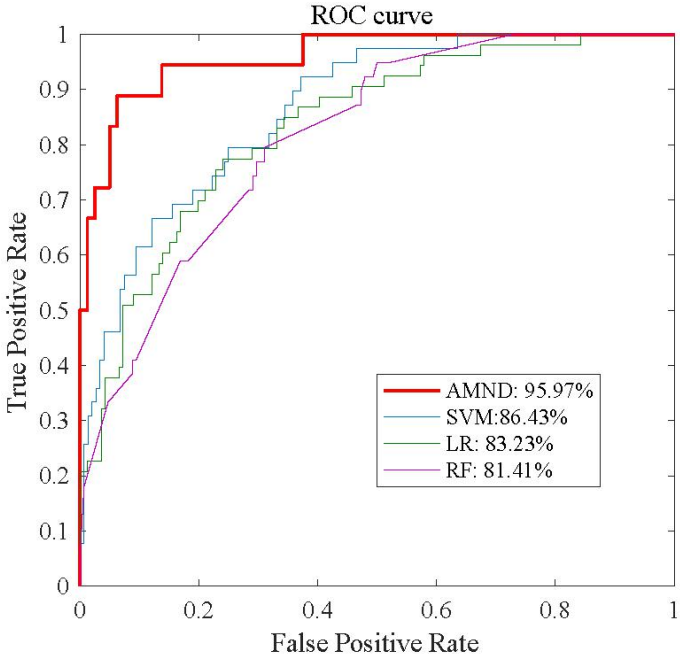

图 4.13　TCGA 数据集上 AMND 算法与 SVM、LR 和 RF 方法的 ROC 曲线

除了 AUC 值，本文还分析了不同方法的 ACC、Pre、F1-score 和 recall 等指标的值，相应的结果如图 4.14 和表 4.5 所示。从图 4.14 可以看出，AMND 方法的整体性能最好，Acc、Pre、F1-score 和 recall 的结果均高出其他三种方法。由表 4.5 可知，AMND 方法在 Acc、Pre、F1-score 和 recall 上分别高出 SVM 方法 10.37%、7.27%、6.62%和 4.49%。此外，AMND 相比于 LR 和 RF 两种方法，同样取得了更

优的性能。总而言之，通过多个性能指标来评估 AMND 方法的性能，可看出 AMND 方法均优于其他几种方法，进一步证实了该方法在乳腺癌生存期预测上的有效性。

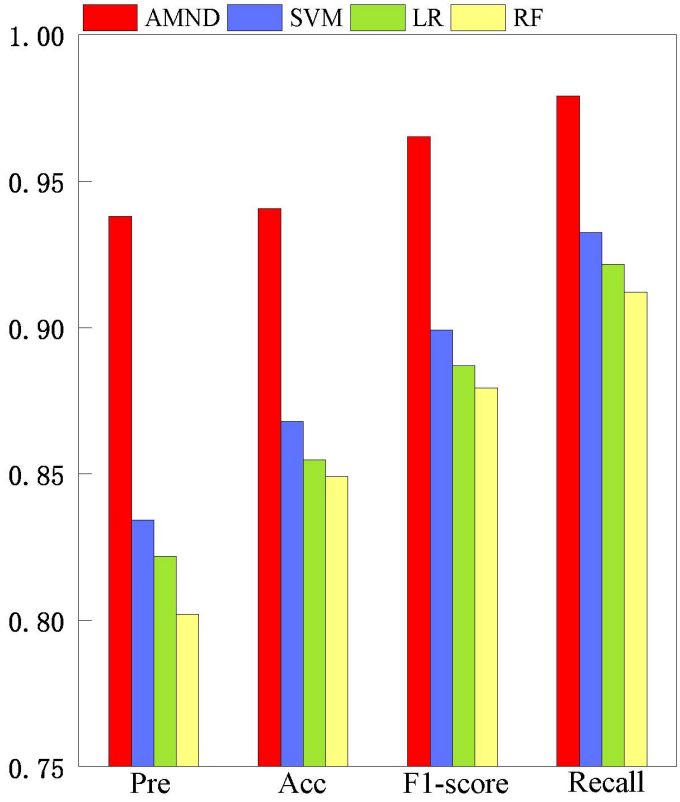

图 4.14　TCGA 数据集上 AMND 算法与 SVM、LR 和 RF 方法的预测性能

表 4.5　TCGA 数据集上 AMND 算法与 SVM、LR 和 RF 方法的预测性能

Method	Acc	Pre	F1-score	recall
SVM	83.42%	86.79%	89.9%	93.42%
LR	82.19%	85.47%	88.69%	92.16%
RF	80.21%	84.9%	87.94%	91.21%
AMND	**93.79%**	**94.06%**	**96.52%**	**97.91%**

4.4 本章小结

本章提出了一种基于组学数据和注意力机制的深度神经网络模型（AMND）用于乳腺癌生存期预测。为了高效提取基因表达数据和临床数据中的重要特征，使用 Attention 机制融合 Multi_NMF 算法提取的五种特征向量，得到新的基因表达特征，再将其与临床数据结合，放入神经网络中进行预测。实验结果表明，Attention 机制不仅考虑了患者临床数据和基因表达数据间的关联，还利用注意力机制自适应融合了多种特征提取算法。综上所述，AMND 方法获取了乳腺癌患者的重要组学特征，提升了生存期预测模型的性能。因此，本章的研究对深入了解乳腺癌组

学数据和开发相关生存期预测方法均具有重要意义。

第 5 章 基于多尺度特征融合的生存期预测模型

在第四章的研究中，本文采用深度学习方法融合组学数据，实验结果表明融合后的组学数据能有效提高乳腺癌生存期预测准确率。然而上述方法将融合后的组学数据直接输入神经网络进行预测，具有尺度不变性。这意味着神经网络的感受野是固定的，导致模型很难学到多尺度特征信息。最近，基于多尺度特征融合的深度神经网络能挖掘数据的不同粒度特征，被广泛应用于目标检测和图像处理等领域。例如，Li[67]使用前馈神经网络和多尺度技术建立了一种 TxtNet 模型，该模型捕获了不同尺度的上下文信息，使文本具有更好的语义相关性。受上述研究启发，本文构建了基于多尺度特征融合的深度神经网络模型用于乳腺癌生存期预测。该模型设置了不同大小的池化层对组学数据进行池化运算，使得模型能够挖掘不同尺度的组学特征，从而进一步提升模型的预测性能。

5.1 基于多尺度特征融合的生存期预测模型

本章提出了一种基于多尺度特征融合的深度神经网络模型（Multi-scale Feature Fusion Deep neural networks，MFFD）来进行乳腺癌生存期预测。第四章中的 AMND 方法将融合后的组学数据特征直接输入神经网络，网络的感受野是固定的，无法从多尺度来获取更深层次的组学特征。因此，本文采用多个不同大小的池化操作来获得不同尺度的特征信息。该方法不仅能优化网络结构，还提升了乳腺癌生存期预测准确率。MFFD 的结构图如图 5.1 所示。

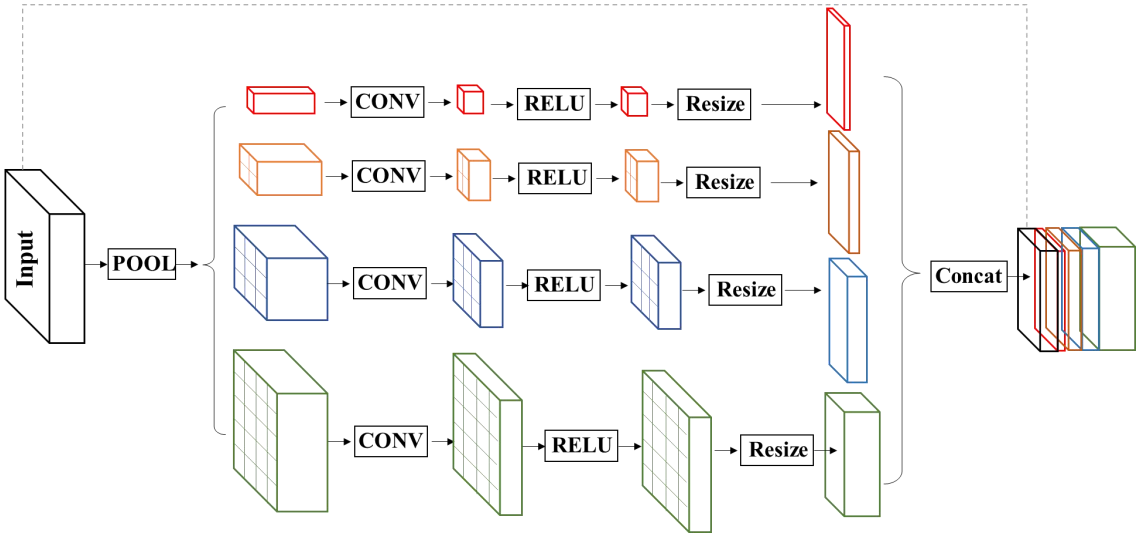

图 5.1 多尺度特征融合的生存期预测模型图

首先，将融合后的组学数据进行 4 种不同大小的池化操作，池化层分别为1×1、2×2、3×3 和 5×5，得到不同尺度的组学特征。其次，为了防止由于多尺度带来的复杂计算，在最大池化后进行1×1 卷积，这样不仅能减少维度，还能进一步提取

局部特征。接着，对卷积后得到的特征使用 Relu 激活函数，从而去掉无关特征。最后，为了获得与最初 Input 相同尺寸，将 Relu 后的特征以线性插值的方法进行 Resize，得到相同的维度，并将不同尺度的特征进行拼接（Concatenate）操作，输入神经网络进行预测。MFFD 方法能从不同的角度提取丰富的局部特征，不仅提高了重要特征的利用率，让模型更加高效，还进一步提升了乳腺癌生存期预测性能。

5.2 实验数据和参数设置

本章的开发环境同样基于谷歌开源的 Tensorflow 深度学习框架和 Python 编程语言。操作系统为 Ubuntu 18.04，GPU 型号为 NVIDIA GeForce GTX 1080，显存 8G，内存 32G，CPU 型号为 Intel(R) Core(TM) i7-7700K CPU@4.20GHz。实验数据依旧选用 METABRIC 数据集和 TCGA 数据集中的基因表达数据和临床数据。

超参数的选择能极大影响模型的性能，在 MFFD 模型的训练过程中，本文使用 Adam 优化器来调节神经网络中的参数，并且激活函数为 Tanh 函数，学习率设置为 0.001，网络中的最小批量尺寸为 32。本章采用最大池化方式，并使用上一章的方式来划分数据集，在测试集上得到模型的最终预测结果。

5.3 实验与结果分析

5.3.1 实验设计

为了验证本章提出的 MFFD 模型的有效性，本章设置对比实验如下：

（1）MFFD 模型与基于不同尺度的深度神经网络模型进行对比。在 MFFD 模型中，本章设计了四种不同大小的尺度，来获取不同粒度的组学特征。为了验证不同尺度对乳腺癌生存期预测的影响，本章设置基于两种、三种和五种尺度大小的模型进行对比。本实验可分析不同尺度的模型对生存期预测性能的影响，同时，选出最佳的多尺度模型。

（2）MFFD 模型与现有的生存期预测方法对比。一方面，本章将 MFFD 方法分别与第三章和第四章所提出的方法进行对比，可验证多尺度特征融合模型能获取深层次组学特征。另一方面，将本章所提出的多尺度特征融合模型与同研究领域内的其他研究人员所提出的生存期预测方法进行对比。本实验通过自我模型对比和与他人工作对比来说明基于多尺度特征融合的生存期预测模型的有效性和可行性。

5.3.2 不同尺度的深度神经网络模型性能分析

MFFD 方法是将融合后组学数据进行 4 种不同大小的池化操作，池化层分别为1×1，2×2，3×3和5×5，得到不同尺度的组学特征，再将这些特征进行融合。

为了验证 MFFD 模型的有效性，本文设置几种不同的池化层参数。例如，融合 2 种池化操作所得的特征，池化层分别为1×1 和 2×2，这里将此模型命名为 MFFD_2。融合 3 种池化操作所得的特征时，池化层分别为1×1，2×2 和 3×3，这里将此模型命名为 MFFD_3。当融合 5 种池化操作所得的特征时，池化层分别为1×1，2×2，3×3，5×5 和10×10，将此模型简称为 MFFD_5。基于不同尺度的深度神经网络模型的实验结果如图 5.2 所示。从图 5.2 中可以看出，不同尺度的模型具有不同的预测效果。MFFD 模型的 AUC 值为 89.24%。MFFD_5 模型、MFFD_3 模型和 MFFD_2 模型的 AUC 值分别为 88.12%、87.75%和 87.67%。MFFD 模型的 AUC 值相较于其他尺度模型的 AUC 值，分别提升了 1.12%、1.49%和 1.57%。进一步说明了融合多尺度特征可以获取不同粒度的特征，进而提升模型预测性能。

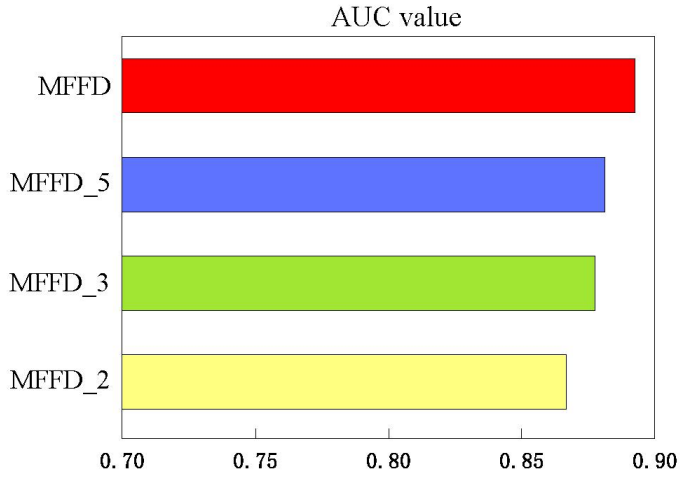

图 5.2　METABRIC 数据集上基于不同尺度的 MFFD 模型的 AUC 值

除了 AUC 值，我们还对比了准确率、精准率、F1 值和召回率等指标。实验结果如表 5.1 所示。

表 5.1　METABRIC 数据集上不同尺度的 MFFD 模型的性能对比

Method	AUC	Acc	Pre	F1-score	recall
MFFD_2	87.67%	83%	83.97%	89.22%	89.75%
MFFD_3	87.75%	85.77%	84.35%	90.8%	92.16%
MFFD_5	88.12%	84.86%	85.63%	90.32%	92.77%
MFFD	**89.24%**	**86.23%**	**89.47 %**	**91.57%**	**98.19%**

从表 5.1 中可以看出，MFFD 方法的整体性能明显优于其他方法。MFFD 方法的精确率为 89.47%，而 MFFD_5、MFFD_3 和 MFFD_2 方法的精确率分别为 85.63%、84.35%和 83.97%。MFFD 方法比 MFFD_5、MFFD_3 和 MFFD_2 方法的精确率分别高 3.84%、5.12%和 5.5%。同时，MFFD 方法的准确率、F1 值和召回率分别为 86.26%、91.57%和 98.19%，相较于 MFFD_5、MFFD_3 和 MFFD_2 方法

也有一定的提升。从实验结果可以看出，MFFD（即取 4 个不同大小的池化层）方法的预测结果比 MFFD_5 好。由此可见，并非尺度越多，预测效果就越好。同时我们可看出基于多尺度特征融合的深度学习方法，在乳腺癌生存期预测中效果显著。

同理，本文将在 TCGA 数据集上进一步验证多尺度特征融合模型的有效性。实验结果如图 5.3 所示。从图 5.3 中可以看出，融合不同池化操作所得的特征，预测效果也不同。就多尺度特征融合模型而言，MFFD 方法的预测效果最好，AUC 值为 98.72%。MFFD_5 方法、MFFD_3 方法和 MFFD_2 方法的 AUC 值分别为 96.98%、97.04%和 97.5%。MFFD 模型的 AUC 值相较于其他尺度模型的 AUC 值，分别提升了 1.74%、1.68%和 1.22%。总体实验结果与 METABRIC 数据集上的结果一致，进一步验证了 MFFD 方法能够获取更好的组学数据特征。

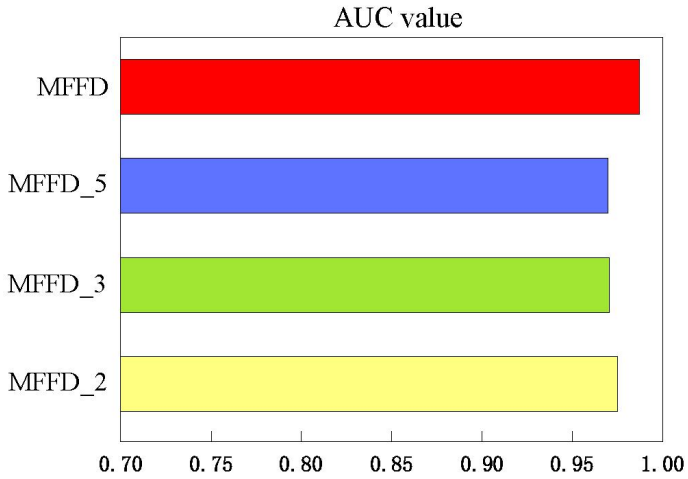

图 5.3　TCGA 数据集上基于不同尺度的 MFFD 模型的 AUC 值

除了 AUC 值，我们还在 TCGA 数据集上对比了准确率、精准率、F1 值和召回率等指标。实验结果如表 5.2 所示。

表 5.2　TCGA 数据集上不同尺度的 MFFD 模型的性能对比

Method	Auc	Acc	Pre	F1-score	recall
MFFD_2	97.5%	95.26%	96.45%	97.14%	95.6%
MFFD_3	97.04%	94.02%	95.12%	97.63%	97.55%
MFFD_5	96.98%	93.96%	94.78%	96.03%	97.32%
MFFD	**98.72%**	**96.47%**	**97.61%**	**98%**	**98.4%**

由表 5.2 可知，MFFD 方法明显优于其他方法。MFFD 方法的准确率、精确率、F1 值和召回率分别为 96.47%、97.61%、98%和 98.4%，相较于 MFFD_5、MFFD_3 和 MFFD_2 方法有明显的提高。由此可见，基于多尺度特征融合的深度神经网络模型能获取不同粒度的组学特征，从而进一步提升模型的预测准确率。

5.3.3 现有方法性能比较

为了进一步检验 MFFD 方法在乳腺癌生存期预测方面的优异表现，将本章提出的 MFFD 模型分与第三章所提出的 Multi_NMF 方法和第四章所提出的 AMND 方法进行对比和分析。图 5.4 展示了 METABRIC 数据集上 MFFD 模型、AMND 模型和 Multi_NMF 模型的 ROC 曲线。从图 5.4 中可以看出，相较于 AMND 方法和 Multi_NMF 方法，MFFD 模型具有更好的预测性能，AUC 值为 89.26%。比第三章提出的 Multi_NMF 方法和第四章提出的 AMND 方法的 AUC 高 2.2%和 10.18%。由此说明，多尺度特征融合可以获得更多与乳腺癌生存期相关的组学特征，从而提升预测效果。

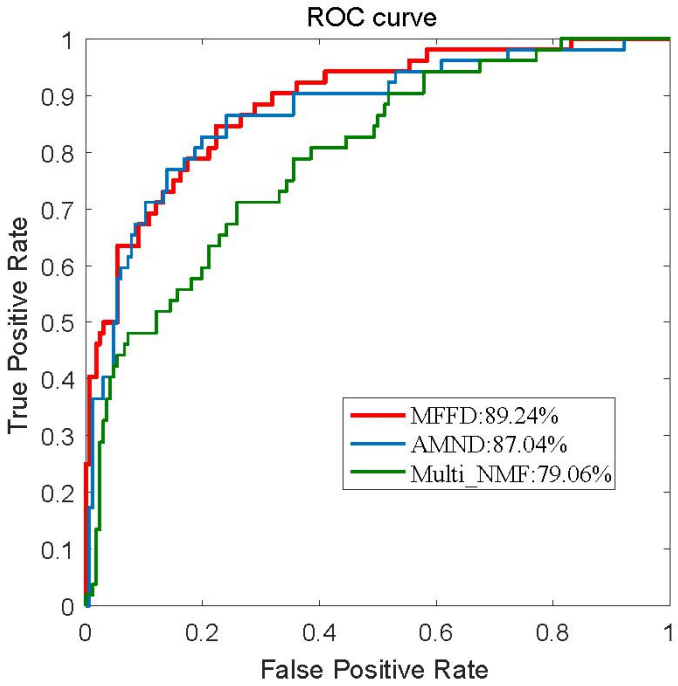

图 5.4　METABRIC 数据集上 MFFD、AMND 和 Multi_NMF 模型的 ROC 曲线

表 5.3　METABRIC 数据集上 MFFD、AMND 和 Multi_NMF 模型的性能对比

方法	AUC	Accuracy	Precision
Multi_NMF	79.06%	80.31%	83.97%
AMND	87.04%	83.13%	84.35%
MFFD	**89.24%**	**86.59%**	**86.39%**

除了 AUC 值，本文还采用其他指标进行对比，如准确率(Accuracy)和精确率（Precision）。实验结果如表 5.3 所示。由表 5.3 可知，MFFD 在整体性能上均有所提升。不仅 AUC 值比 AMND 模型和 Multi_NMF 模型有所提高，同时准确率和精确率也高于 AMND 和 Multi_NMF。MFFD 模型的预测准确率为 86.59%，分别比

AMND 模型和 Multi_NMF 模型高 3.46%和 6.28%。MFFD 模型的预测精确率为 86.59%，分别比 AMND 模型和 Multi_NMF 模型高 2.04%和 2.24%。

同理，我们将进一步分析 TCGA 数据集上 MFFD 方法与第三章所提出的 Multi_NMF 方法和第四章所提出的 AMND 方法的实验结果。图 5.5 展示了 TCGA 数据集上 MFFD 模型、AMND 模型和 Multi_NMF 模型的 ROC 曲线示意图。对比实验结果可知，相较于 AMND 方法和 Multi_NMF 方法，MFFD 模型具有更好的预测性能，AUC 值为 98.72%。比第三章提出的 Multi_NMF 方法和第四章提出的 AMND 方法的 AUC 高 2.75%和 13.45%。由此说明，多尺度特征融合可以获得更多与乳腺癌生存期相关的特征，从而提升预测效果。

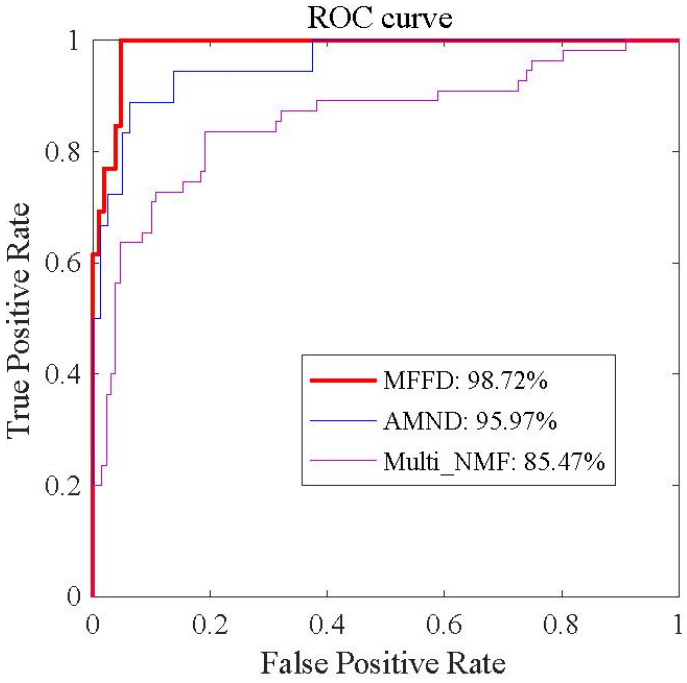

图 5.5　TCGA 数据集上 MFFD、AMND 和 Multi_NMF 模型的 ROC 曲线

表 5.4　TCGA 数据集上 MFFD、AMND 和 Multi_NMF 模型的性能对比

方法	AUC	Accuracy	Precision
Multi_NMF	85.47%	84.64%	86.56%
AMND	95.97%	93.79%	94.06%
MFFD	**98.72%**	**96.47%**	**97.61%**

除了 AUC 值，本文还采用其他指标进行对比，如准确率(Accuracy)和精确率（Precision）。实验结果如表所示。由表 5.4 可知，MFFD 在整体性能上均有所提升。不仅 AUC 值比 AMND 模型和 Multi_NMF 模型有所提高，同时准确率和精确率也高于 AMND 和 Multi_NMF。MFFD 模型的预测准确率为 96.47%，分别比 AMND 模型和 Multi_NMF 模型高 2.68%和 11.83%。而且 MFFD 模型的预测精确

分别比 AMND 模型和 Multi_NMF 模型高 3.55%和 11.05%。综上所述，基于多尺度特征融合的方法确实能得到更具有生物意义的特征，从而提高乳腺癌生存期预测效果。

为了验证 MFFD 模型的性能优劣，本文将 MFFD 方法与其他乳腺癌生存期预测研究的结果进行对比。例如，Sun 等人[66]针对 METABRIC 数据中的基因表达谱、CAN 谱和临床数据进行生存期预测研究，提出了 MDNNMD 方法，得到的 AUC 值为 84.5%。Gevaert[23]等人搭建了一种基于贝叶斯网络的概率图模型（BPIM）来融合乳腺癌患者的临床数据和基因表达数据，在独立测试集上取得了 84.5%的 AUC 值。Khademi 等人[64]提出了概率图模型（PGM）融合 METABRIC 数据中的基因表达谱和临床数据，得到的 AUC 值为 82%。如表 4.5 所示，MFFD 方法分别比 MDNNMD、PGM 和 BPIM 的 AUC 值分别高 4.74%、4.74%和 7.24%。由此可见，MFFD 方法在乳腺癌生存期预测上取得了很好的效果，进一步证实了融合组学数据的深度学习方法对乳腺癌生存期预测的可行性和有效性。

表 5.5　METABRIC 数据集上 MFFD 和现有研究成果比较

Measure	AUC
MDNNMD	84.5%
BPIM	84.5%
PGM	82%
MFFD	**89.24%**

同理，本文将进一步在 TCGA 数据集上对比 MFFD 方法和其他乳腺癌生存期预测研究的结果。

表 5.6　TCGA 数据集上 MFFD 和现有研究成果比较

Measure	AUC
MDNNMD	93.8%
文献[5]	83.93%
文献[68]	82%
MFFD	**98.72%**

例如，Sun 等人[66]提出的 MDNNMN 方法在 TCGA 数据集上的 AUC 值为 93.8%。文献[5]通过融合 DNA 甲基化、拷贝数变异、蛋白质表达和基因表达四种组学数据，来预测乳腺癌患者的生存期，这种结合随机森林和组学数据的算法取得了 83.93%的 AUC 值。文献[68]结合基于通路失调评分（PDS）的通路算法、Cox 回归和 L1-LASSO 惩罚方法，开发了一种新型的乳腺癌预后计算模型，且该模型在 TCGA 数据集上的 AUC 值为 80%。表 5.6 展示了 MFFD 方法与上述研究的 AUC 值。由表 5.6 可知，MFFD 方法比上述研究的预测效果更好，AUC 值为 98.72%。

通过上述分析我们可知：基于深度学习的组学数据融合方法比传统的机器学习方法能更好的预测乳腺癌生存期。

5.4 本章小结

本章提出了一种基于多尺度特征融合的生存期预测模型（MFFD）。在神经网络中，固定感受野输出的特征具有尺度不变性。为了得到组学数据不同粒度的特征，本文设计大小不同的池化层，来获取不同尺度的组学特征。通过详细对比 MFFD 方法与当前乳腺癌生存期预测方法，实验结果表明，MFFD 方法在乳腺癌生存期预测中具有更好的预测能力。因此，基于多尺度特征融合的深度神经网络模型能挖掘更多的组学特征，具有更丰富的生物学意义。

第 6 章 总结与展望

6.1 总结

乳腺癌是危害女性生命健康的恶性肿瘤，乳腺癌生存期作为乳腺癌预后研究的重要指标，对癌症病人的心理康复和临床治疗具有重要意义。现有的乳腺癌生存期预测方法存在使用单一特征提取方法、未考虑组学数据等问题，导致难以获取组学数据中的深层次特征。针对当前研究存在的缺陷，本文提出了基于组学数据和深度学习的乳腺癌生存期预测方法。本文整体的研究工作归纳如下：

（1）基于现有的非负矩阵分解算法，提出一种 Multi_NMF 特征提取算法。改进后的 Multi_NMF 算法平衡了特征间的相关性和冗余性，一定程度上改善了重要特征丢失的问题。在 METABRIC 数据集和 TCGA 数据集上验证 Multi_NMF 方法的有效性，实验结果表明，改进后的 Multi_NMF 方法具有较好的特征提取能力，同时取得了更好的预测效果。

（2）为了获取组学数据中更全面的生物学信息，本文提出了一种基于组学数据和注意力机制的生存期预测模型（AMND）。首先，该方法使用五种不同的非负矩阵分解优化算法对基因表达数据进行特征提取，以获取更全面的特征。其次，基于 Attention 机制的组学数据融合方法考虑了组学数据间的相关性，能够提取基因表达数据和临床数据中隐含的生物特征。最后，实验结果一致表明，该方法比现有的乳腺癌生存期预测方法取得了更高的预测准确率。

（3）以上述研究为基础，本文构建了一种基于多尺度特征融合的生存期预测模型（MFFD）。该方法将组学数据输入不同大小的池化层，以获取不同尺度的组学特征，并将不同尺度的特征进行融合。基于多尺度特征融合的深度神经网络模型能够全方位的捕获组学数据中与生存期密切相关的特征信息，同时使网络具有较高的敏感性和特异性。将 MFFD 方法与现有的癌症生存期预测方法对比，结果表明，MFFD 能够更准确地预测乳腺癌生存期。

本文研究了基于深度学习的乳腺癌生存期预测方法，通过自我模型对比以及与同研究领域内其他研究方法对比，实验结果表明，本文提出的乳腺癌生存期预测方法具有较好的预测效果。但由于时间和条件的现实，还存在很多问题和难点需要进一步探究。

6.2 展望

综上所述，本文的研究取得了较好的实验效果，为推进乳腺癌生存期预测研究做出了微小贡献。然而，该方法也存在一些不足之处，有待在未来的研究中进一步完善。将来的研究工作可以从以下三个方面进行：

（1）本文提出的基于深度学习和组学数据融合的乳腺癌生存期预测方法，虽然取得了较好的预测效果，但仍存在完善的空间。随着病理图像的增多，考虑更多模态的生物学数据，如乳腺癌患者的病理图像和其他类型的组学数据，获取多模态数据特征的同时进一步提升癌症生存期预测准确率，是未来研究的重点内容。

（2）本文实验基于 METABRIC 数据集和 TCGA 数据集，但深度学习方法对于数据量有极高要求。本文的数据集包含的样本数量较少，可能会影响到网络模型的预测效果。受益于高精度的基因检测技术，越来越多的乳腺癌数据集将得到扩大和补充。基于大数据的乳腺癌生存期预测具有更强的可解释性。

（3）在多尺度特征融合模型中，多个池化层增加了模型的复杂度，实验的用时过长，实验效率有待提高。因此，在未来的工作中，应进一步优化模型结构，在提升预测性能的同时降低模型的复杂性。

参考文献

[1] Siegel R L, Miller K D, Jemal A J C a C J F C. Cancer statistics, 2016[J], 2016, 66(1): 7-30.

[2] Honein-Abouhaidar G N, Hoch J S, Dobrow M, et al. Cost analysis of breast cancer diagnostic assessment programs[J], 2017, 24(5): e354.

[3] Dai X, Cheng H, Bai Z, et al. Breast cancer cell line classification and its relevance with breast tumor subtyping[J], 2017, 8(16): 3131.

[4] Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J], 2018, 68(6): 394-424.

[5] 齐惠颖,基于多组学数据融合构建乳腺癌生存预测模型[J], 数据分析与知识发现, 2019, 3(8): 88-93.

[6] Hanahan D, Weinberg R a J C. Hallmarks of cancer: the next generation[J], 2011, 144(5): 646-674.

[7] Peng C, Li A J I a T O C B, Bioinformatics. A heterogeneous network based method for identifying GBM-related genes by integrating multi-dimensional data[J], 2016, 14(3): 713-720.

[8] 刘伟, 朱云平,系统生物学研究中不同组学数据的整合[J], 中国生物化学与分子生物学报, 2007, 23(12): 971-976.

[9] Etcheverry A, Aubry M, De Tayrac M, et al. DNA methylation in glioblastoma: impact on gene expression and clinical outcome[J], 2010, 11(1): 701.

[10] Van De Vijver M J, He Y D, Van't Veer L J, et al. A gene-expression signature as a predictor of survival in breast cancer[J], 2002, 347(25): 1999-2009.

[11] Glas A M, Floore A, Delahaye L J, et al. Converting a breast cancer microarray signature into a high-throughput diagnostic test[J], 2006, 7(1): 278.

[12] Xu X, Zhang Y, Zou L, et al. A gene signature for breast cancer prognosis using support vector machine[C]. 2012 5th International Conference on BioMedical Engineering and Informatics, 2012: 928-931.

[13] Stirzaker C, Zotenko E, Song J Z, et al. Methylome sequencing in triple-negative breast cancer reveals distinct methylation clusters with prognostic value[J], 2015, 6: 5899.

[14] Volinia S, Croce C M J P O T N a O S. Prognostic microRNA/mRNA signature

from the integrated analysis of patients with invasive breast cancer[J], 2013, 110(18): 7413-7417.

[15] Wu Y, Chen H, Jiang G, et al. Genome-wide association study (GWAS) of germline copy number variations (CNVs) reveal genetic risks of prostate cancer in chinese population[J], 2018, 9(5): 923.

[16] Mariette J, Villa-Vialaneix N J B. Unsupervised multiple kernel learning for heterogeneous data integration[J], 2017, 34(6): 1009-1015.

[17] Kim D, Li R, Lucas A, et al. Using knowledge-driven genomic interactions for multi-omics data analysis: metadimensional models for predicting clinical outcomes in ovarian carcinoma[J], 2016, 24(3): 577-587.

[18] Ahmad A, Fröhlich H J B. Towards clinically more relevant dissection of patient heterogeneity via survival-based Bayesian clustering[J], 2017, 33(22): 3558-3566.

[19] Coretto P, Serra A, Tagliaferri R J B. Robust clustering of noisy high-dimensional gene expression data for patients subtyping[J], 2018, 34(23): 4064-4072.

[20] Min S, Lee B, Yoon S J B I B. Deep learning in bioinformatics[J], 2017, 18(5): 851-869.

[21] Chaudhary K, Poirion O B, Lu L, et al. Deep learning–based multi-omics integration robustly predicts survival in liver cancer[J], 2018, 24(6): 1248-1259.

[22] Chai H, Zhou X, Cui Z, et al. Integrating multi-omics data with deep learning for predicting cancer prognosis[J], 2019: 807214.

[23] Gevaert O, Smet F D, Timmerman D, et al. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks[J], 2006, 22(14): e184-e190.

[24] Sun Y, Goodison S, Li J, et al. Improved breast cancer prognosis through the combination of clinical and genetic markers[J], 2006, 23(1): 30-37.

[25] Bahdanau D, Cho K, Bengio Y J a P A. Neural machine translation by jointly learning to align and translate[J], 2014.

[26] Luong M-T, Pham H, Manning C D J a P A. Effective approaches to attention-based neural machine translation[J], 2015.

[27] Mnih V, Heess N, Graves A. Recurrent models of visual attention[C]. Advances in neural information processing systems, 2014: 2204-2212.

[28] Zhang Q, Gong Y, Wu J, et al. Retweet prediction with attention-based deep neural network[C]. Proceedings of the 25th ACM international on conference on information and knowledge management, 2016: 75-84.

[29] Chen Y, Zhu X, Gong S, et al. Person Re-Identification by Deep Learning Multi-Scale Representations[C]. 2017 IEEE International Conference on Computer Vision Workshop (ICCVW), 2017.

[30] Carene D, Tran-Dien A, Lemonnier J, et al. Association between FGFR1 copy numbers, MAP3K1 mutations, and survival in axillary node-positive, hormone receptor-positive, and HER2-negative early breast cancer in the PACS04 and METABRIC studies[J], 2019: 1-15.

[31] Zhang X, Zhang W, Jiang Y, et al. Identification of functional lncRNAs in gastric cancer by integrative analysis of GEO and TCGA data[J], 2019.

[32] Liu A, Zhang S, Shen Y, et al. Association of mRNA expression levels of Cullin family members with prognosis in breast cancer: An online database analysis[J], 2019, 98(31).

[33] Yang J, Cai H, Xiao Z-X, et al. Effect of radiotherapy on the survival of cervical cancer patients: An analysis based on SEER database[J], 2019, 98(30).

[34] Hiasa Y, Otake Y, Takao M, et al. Automated Muscle Segmentation from Clinical CT using Bayesian U-Net for Personalized Musculoskeletal Modeling[J], 2019.

[35] Curtis C, Shah S P, Chin S-F, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups[J], 2012, 486(7403): 346.

[36] 蔡瑞初, 侯永杰,基于层级规则树的跨平台基因表达数据分类[J], 计算机工程, 2019(7): 5.

[37] Speed T. Statistical analysis of gene expression microarray data[M]. Chapman and Hall/CRC, 2003.

[38] Donoho D L J a M C L. High-dimensional data analysis: The curses and blessings of dimensionality[J], 2000, 1(2000): 32.

[39] 谢雁鸣,纵向数据分析方法在中医临床疗效评价中的应用浅析[J], 中国中医基础医学杂志, 2007, 13(9): 711-713.

[40] Liang K-Y, Zeger S L J B. Longitudinal data analysis using generalized linear models[J], 1986, 73(1): 13-22.

[41] Abreu P H, Amaro H, Silva D C, et al. Overall survival prediction for women breast cancer using ensemble methods and incomplete clinical data[C]. XIII Mediterranean Conference on Medical and Biological Engineering and Computing 2013, 2014: 1366-1369.

[42] Lecun Y, Bengio Y, Hinton G J N. Deep learning[J], 2015, 521(7553): 436-444.

[43] Hagan M T, Menhaj M B J I T O N N. Training feedforward networks with the

Marquardt algorithm[J], 1994, 5(6): 989-993.
[44] Bottou L: Large-scale machine learning with stochastic gradient descent, Proceedings of COMPSTAT'2010: Springer, 2010: 177-186.
[45] Konečný J, Liu J, Richtárik P, et al. Mini-batch semi-stochastic gradient descent in the proximal setting[J], 2015, 10(2): 242-255.
[46] Goodfellow I, Bengio Y, Courville A J C, Massachusetts. Deep Learning| The MIT Press[J], 2016.
[47] Gu J, Wang Z, Kuen J, et al. Recent advances in convolutional neural networks[J], 2018, 77: 354-377.
[48] Lecun Y: Generalization and network design strategies, Connectionism in perspective: Citeseer, 1989.
[49] Lee D D, Seung H S J N. Learning the parts of objects by non-negative matrix factorization[J], 1999, 401(6755): 788.
[50] Vavasis S a J S J O O. On the complexity of nonnegative matrix factorization[J], 2009, 20(3): 1364-1377.
[51] Lee D D, Seung H S. Algorithms for non-negative matrix factorization[C]. Advances in neural information processing systems, 2001: 556-562.
[52] Paatero P, Tapper U J E. Positive matrix factorization: A non‐negative factor model with optimal utilization of error estimates of data values[J], 1994, 5(2): 111-126.
[53] Baktash H, Natale E, Viennot L J a P A. A Comparative Study of Neural Network Compression[J], 2019.
[54] Koiran P, Sontag E D. Neural networks with quadratic VC dimension[C]. Advances in neural information processing systems, 1996: 197-203.
[55] Lin C-J J N C. Projected gradient methods for nonnegative matrix factorization[J], 2007, 19(10): 2756-2779.
[56] Bayar B, Bouaynaya N, Shterenberg R J J O B, et al. Probabilistic non-negative matrix factorization: theory and application to microarray data analysis[J], 2014, 12(01): 1450001.
[57] 王晓丹, 雷蕾,基于混淆矩阵的证据可靠性评估[J], 系统工程与电子技术, 2016, 37(4).
[58] 张颖. 基于乳腺癌基因表达数据的特征选择算法研究[D]. 西南大学, 2019.
[59] Yuvaraj N, Vivekanandan P. An efficient SVM based tumor classification with symmetry non-negative matrix factorization using gene expression data[C]. 2013

International Conference on Information Communication and Embedded Systems (Icices), 2013: 761-768.

[60] Jazayeri N, Sajedi H J S a S. Breast cancer diagnosis based on genomic data and extreme learning machine[J], 2020, 2(1): 3.

[61] Liu X, Shi J, Wang C. Hessian regularization based non-negative matrix factorization for gene expression data clustering[C]. 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015: 4130-4133.

[62] Ding Z, Zu S, Gu J J B. Evaluating the molecule-based prediction of clinical drug responses in cancer[J], 2016, 32(19): 2891-2895.

[63] Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays[J], 2001, 17(6): 520-525.

[64] Khademi M, Nedialkov N S. Probabilistic graphical models and deep belief networks for prognosis of breast cancer[C]. 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), 2015: 727-732.

[65] Simsek S, Kursuncu U, Kibis E, et al. A hybrid data mining approach for identifying the temporal effects of variables associated with breast cancer survival[J], 2020, 139: 112863.

[66] Sun D, Wang M, Li A J I a T O C B, et al. A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data[J], 2018, 16(3): 841-850.

[67] Li C, Deng C, Li N, et al. Self-Supervised Adversarial Hashing Networks for Cross-Modal Retrieval[J].,2016.

[68] Huang S, Yee C, Ching T, et al. A novel model to combine clinical and pathway-based transcriptomic information for the prognosis prediction of breast cancer[J], 2014, 10(9).

攻读硕士期间发表论文及科研工作

发表的学术论文：

1. Attention-Based Multi-NMF Deep Neural Network with Multimodality Data for Breast Cancer Prognosis Model[J]
2. Face Recognition Algorithm Based on VGG Network Model and SVM[C]