

**毕业设计（论文）**

题 目 基于肠道微生物组的

疾病辅助诊断系统

专 业 计算机科学与技术

学　　 号 180400719

学 生 张博玮

指 导 教 师 徐永东

答 辩 日 期 2022年6月7 日

# 摘 要

据现有相关学术分析，某些特定的疾病与寄宿在人体内的微生物群落丰度息息相关，而人体内超过80%的微生物都寄宿在肠道之内，作为人体内最大的微生物寄宿场所，肠道中的微生物组对人体健康和疾病有着直接的影响。近年来高通量测序技术16SrDNA与PCR技术在碱基测序并根据测序结果进行多样性分析的领域已经取得了广泛的应用，以上两种技术的兴起提供了真实可信的肠道内微生物丰度数据，这为后续进行的基于机器学习的方法分析肠道微生物群落的特征以及与其他领域的关联关系提供了诸多可行方案，因此本研究将机器学习方法应用于微生物与疾病关联关系分析领域，并基于本研究搭建了一个供医生使用的辅助诊断系统。

本研究共囊括2424个肠道微生物丰度测序样本，源自于八项研究六种疾病，测序数据来自于鸟枪法宏基因组学。本研究首先应用支持向量机模型(Support Vector Machine, SVM)和随机森林模型(Random Forest, RF)来寻找这些样本与疾病之间的关联关系，以根据这些关联关系判断样本患病或者健康，即进行样本分类任务。机器学习方法最终取得了较好的结果，其中随机森林模型在所有疾病上取得的accuracy和auc都高于支持向量机模型，这说明随机森林算法更契合本研究。随后本研究搭建了全连接神经网络模型和一维卷积神经网络模型来完成样本分类任务，在两个数据集上取得了超越随机森林模型的accuracy。此外，本研究将图神经网络应用于疾病分类任务，图神经网络在一个数据集上的accuracy超越了随机森林模型，但是总体结果不如以上两种深度学习模型。从研究结果上看，表现最好的随机森林模型可以在cirrhosis数据集上得到0.877的accuracy和0.945的auc。其余方法能取得的最佳accuracy都是在cirrhosis数据集上，结果介于0.799到0.853之间。以上方法在colorectal数据集和IBD数据集上取得的结果稍逊色于cirrhosis数据集，而在剩余三个数据集上表现较差，accuracy普遍介于0.593于0.703之间。随后本研究将上述方法应用在了真实的疾病诊断场景，可以诊断操作员上传的OTU表中的样本为患病或者健康，如果样本患病，还可以采用逆因果算法给出微生物丰度层面的诊疗建议以作为疾病辅助诊断系统，该诊断环节中的结果也在其他的研究中被加以印证。

关键词：图神经网络；机器学习；深度学习；肠道菌群；疾病诊断；

# Abstract

According to existing academic analyses, certain diseases are closely related to the abundance of microbial communities residing in the human body, and more than 80% of the microorganisms in the human body reside in the gut, which is the largest microbial host site in the human body, and the microbiome in the gut has a direct impact on human health and diseases. In recent years, high-throughput sequencing 16SrDNA and PCR technologies have been widely used in the field of base sequencing and diversity analysis based on the sequencing results, and the rise of these two technologies has provided real and reliable data on microbial abundance in the gut, which provides many feasible solutions for subsequent machine learning-based methods to analyze the characteristics of the gut microbial community and its association with other fields. Therefore, this study applies machine learning methods to the analysis of microbial-disease associations, and builds a diagnostic aid for physicians based on this study.

A total of 2424 gut microbial abundance sequencing samples from eight studies with six diseases were included in this study, and the sequencing data were obtained from birdshot macrogenomics. This study first applied Support Vector Machine (SVM) and Random Forest (RF) models to find the association between these samples and diseases in order to determine whether the samples are sick or healthy based on these associations, i.e., to perform the sample classification task. The machine learning methods eventually achieved better results, where the Random Forest model achieved higher accuracy and auc than the Support Vector Machine model on all diseases, which indicates that the Random Forest algorithm is a better fit for this study. In addition, this study applied the graph neural network to the disease classification task, and the graph neural network surpassed the random forest model in terms of accuracy on one dataset, but the overall results were not as good as the above two deep learning models. The best performance of the random forest model was 0.877 accuracy and 0.945 auc on the cirrhosis dataset, and the best accuracies achieved by the other methods were on the cirrhosis dataset, with results ranging from 0.799 to 0.853. The above methods achieved slightly worse results on the colorectal and IBD datasets than on the cirrhosis dataset, and performed poorly on the remaining three datasets, with accuracies generally ranging between 0.593 and 0.703. This study then applied the above method to a real disease diagnosis scenario, where the samples in the OTU table uploaded by the operator could be diagnosed as sick or healthy, and if the samples were sick, the inverse causality algorithm could be used to give microbial abundance level treatment recommendations as a disease aid diagnosis system.

**Keywords:** Graph Neural Network, Machine learning, Gut microbe, Disease prediction, Auxiliary Diagnosis

目录

[摘 要 I](#_Toc105055315)

[Abstract II](#_Toc105055316)

[**第1章 绪 论** 1](#_Toc105055317)

[1.1 课题背景及研究的目的和意义 1](#_Toc105055318)

[1.1.1 课题背景 1](#_Toc105055319)

[1.1.2 研究目的和意义 1](#_Toc105055320)

[1.2 国内外研究现状及分析 2](#_Toc105055321)

[1.3 研究内容及关键问题 4](#_Toc105055322)

[1.3.1研究内容： 4](#_Toc105055323)

[1.3.2关键问题： 4](#_Toc105055324)

[1.4 本文组织结构 5](#_Toc105055325)

[**第2章 基于机器学习的肠道微生物与疾病关联关系分析方法** 6](#_Toc105055326)

[2.1数据集来源与构建 6](#_Toc105055327)

[2.1.1 数据来源 6](#_Toc105055328)

[2.1.2 数据预处理 6](#_Toc105055329)

[2.2研究方法 8](#_Toc105055330)

[2.2.1 机器学习方法 8](#_Toc105055331)

[2.2.2 深度学习方法 9](#_Toc105055332)

[2.3 结果分析： 17](#_Toc105055333)

[2.4 本章小结 21](#_Toc105055334)

[**第3章 基于随机森林算法的关键物种正常丰度估计算法** 22](#_Toc105055335)

[3.1 前言 22](#_Toc105055336)

[3.2 算法设计 22](#_Toc105055337)

[3.2 结果与分析 24](#_Toc105055338)

[3.3 本章小结 25](#_Toc105055339)

[**第4章 基于机器学习与肠道微生物组的疾病辅助诊断系统的设计与实现** 26](#_Toc105055340)

[4.1 需求分析 26](#_Toc105055341)

[4.1.1 系统的业务流程 26](#_Toc105055342)

[4.1.2 系统的具体需求 27](#_Toc105055343)

[4.1.3 用例分析 27](#_Toc105055344)

[4.2 概要设计 30](#_Toc105055345)

[4.2.1 功能模块划分 30](#_Toc105055346)

[4.2.2 数据库设计 31](#_Toc105055347)

[4.3 详细设计 33](#_Toc105055348)

[4.4 系统的实现与测试 34](#_Toc105055349)

[4.5 本章小结 43](#_Toc105055350)

[**结 论** 44](#_Toc105055351)

[**本科毕业论文（设计）原创性声明** 45](#_Toc105055352)

[**参考文献** 46](#_Toc105055353)

[**致 谢** 48](#_Toc105055354)

**第1章 绪 论**

## 课题背景及研究的目的和意义

### 1.1.1 课题背景

人体内寄宿着大量的微生物，平均每个人体内寄宿着至少100万亿个微生物，包括真菌、病毒、细菌等，微生物与人体之间一直存在着相互作用的关系，一部分寄宿于人类的微生物是与人类共生的，也就是说这部分微生物不会对我们的健康造成伤害，也有一些益生菌与人体有一种互利关系，还有一些微生物损伤人类健康，越来越多的临床研究表明，特定的微生物含量异常与人类各种疾病的发生密切相关[1,2]，微生物组成的偏差会对宿主的组织造成不同程度的损害，并进一步诱发各种疾病[3]。据现有相关学术分析，某些特定的疾病与寄宿在人体内的微生物群落丰度息息相关[4]，而人体内超过80%的微生物都寄宿在肠道之内，作为人体内最大的微生物寄宿场所，肠道中的微生物组对人体健康和疾病有着直接的影响，肠道微生物丰度以及多样性与多种疾病相关，如肥胖、营养不良和糖尿病等代谢性疾病，炎症性肠病、溃疡性结肠炎和克罗恩病等慢性肠道感染性疾病以及结肠癌和肝癌等恶性肿瘤[5]，近年来随着16SrRNA测序技术的兴起,新的研究表明，肠易激综合征患者肠道中的乳酸菌数量明显下降,而便秘型肠易激综合征患者韦永球菌数量上升，这进一步证实了肠道微生物与人体健康之间存在深刻的联系[6]。

近年来，随着生物信息学的蓬勃发展，人们在微生物的测序领域取得了巨大突破，其中高通量测序技术16SrDNA与PCR技术在碱基测序并根据测序结果进行多样性分析的领域已经取得了广泛的应用，借由上述两种技术的兴起以及大量相关研究的推进，目前在肠道微生物菌群领域，基于以上技术的基因芯片技术以及实时荧光定量PCR技术已经成功应用于肠道微生物群落[7]，并提供了真实可信的肠道内微生物丰度数据，这为后续进行的基于机器学习的方法分析肠道微生物群落的特征以及与其他领域的关联关系提供了诸多可行方案。

### 1.1.2 研究目的和意义

随着宏基因组学，转录组学的不断革新，人们逐渐发现肠道微生物菌群的代谢可以在一定程度上参与人体肠道乃至其他部位的调节，能量循环以及新陈代谢，从这个角度出发，肠道菌群对人体影响深远。正常情况下，人体与寄宿在人体肠道内的微生物之间存在着动态平衡关系，无论是共生、益生或是寄生，微生物与人体之间都借助彼此完成自身所不能单独完成的生物学任务，比如，宿主借助微生物来消化无法直接被降解的食物。

现阶段的各种研究已经揭示了肠道微生物群落相对丰度或者说多样性与疾病之间存在的关联，经历了长久的进化以来，人体内的微生物群落间的相对丰度已经达到了某种平衡[8]。微生物群落中的所有菌株种类都贡献于丰度动态平衡，某种微生物群落的相对丰度的过度改变将会使得肠道微生物与宿主之间平衡失衡，群落之间继而通过能量吸收、胆碱、内毒素血政和脑肠轴等多种途径影响宿主的健康[9] 。

综上所述，肠道微生物的丰度很大程度上主导着疾病的发生，因此该领域有着广阔的医学以及生物学研究前景。大量由肠道微生物群落导致的疾病的预防以及治愈，在社会经济快速发展的当下，越来越成为人们迫切需要解决的健康性需求，相关领域医学技术的发展也赫然成为了追逐健康时代下的“刚需”。然而肠道微生物群落极其庞大，不进行预测直接对疾病和群落之间的关系进行分析将会消耗难以承受的人力和精力，关联预测的技术在二者关系的分析领域是不可或缺的。而随着测序技术的蓬勃发展，大量的数据可以用于机器学习的方法，因此使用机器学习或者深度学习的方法，对现存的大量样本进行预测处理，可以为医学分析以及实践提供研究导向。

## 国内外研究现状及分析

Dlvya Sharma等人提出了一个新的深度学习框架 "phyLoSTM"，该框架结合卷积神经网络和长短期记忆网络（LSTM）进行特征提取，同时分析了纵向微生物组测序数据的时间依赖性，以及用于疾病预测的宿主环境因素。Dlvya Sharma等人模拟了100个带有时间特性的数据集进行模型测试，且在两个真实的纵向人类微生物组数据中测试了该方法，与随机森林相比，模拟数据的AUC为0.897（增加了5%），两组真实数据的AUC分别为0.762（增加了19%）和0.713（增加了8%）[10]。

Pasolli等人采用支持向量机（SVM）和随机森林（RF）模型对微生物-疾病关联进行预测。在这项研究中肝硬化是预测效果最好的疾病，AUC能够达到0.945，其次是IBD和结直肠癌，AUC分别可以达到0.890和0.873。Pasolli等人的研究结果归纳后显示，在所有数据集中随机森林模型中的表现都能够超过SVM模型，因此在基于微生物组丰度数据进行微生物-疾病关联预测时随机森林模型的表现最佳。Pasolli等人使用的模型在跨数据集泛化上的表现平庸[11] 。

Lshan等人使用粪便采样获取的16S宏基因组学数据，通过5种机器学习方法针对729例IBD患者和700例健康者进行分析，在鉴定了差异明显的50种微生物后，通过随机森林模型获得了0.8的AUC；该研究还尝试区分克罗恩病和溃疡性结肠炎这两种IBD亚种，在331例克罗恩病患者和141例溃疡性结肠炎患者上使用随机森林模型，AUC可以达到0.92[12] 。

Yahui Long等人通过研究现存的微生物与疾病的关联图，使用基于图注意力网络(Graph Attention Networks)的方式，来进行疾病预测。首先使用编码器来将现有的微生物-疾病关联压缩为特征，再使用半监督学习方式的解码器来将对特征进行新的微生物-疾病关联评分，同时利用该关联评分重新构建网络，该方法取得了极好的疾病预测效果，在基于HMDAD数据集的测试中，他们提出的 GATMDA 模型实现了最佳性能，平均 AUC 为 0.9554±0.0184，平均 AUPR 为 0.9334±0.0417，分别比第二好的模型6.37%和2.72%[13]。这一模型取得的优异成绩证明基于图机器学习的方法可以在微生物-疾病关联预测领域取得全新水平的优异结果，进一步说明了基于图的机器学习方法所揭示的微生物或者疾病内部关联对于疾病预测而言是一种机器有效的工具和手段。

陈兴等人基于具有相似功能的微生物会协同参与同一类疾病这一机制开发了一个新的计算模型，通过结合基于邻居的协同过滤和基于图的评分方法来预测微生物-疾病的关联分数，并且将分数大于某一阈值的关系对视为关联，也就是该微生物可能会导致该疾病，该方法的核心思想在于同时考虑了微生物组间基于高斯核的相似性和基于症状的疾病相似性，同时该预测方法在HMDAD数据集上使用五折交叉验证and 留一交叉验证方法分别取得了0.9023±0.0031 和 0.9111的AUCs[14].

赵燕等人以疾病与微生物的高斯交互剖面核相似度、疾病症状相似度为依据，结合矩阵分解与标签传播计算微生物-疾病关联得分，大于阈值的得分被视为关联，也就是微生物导致了该疾病，在HMDAD数据集上同时通过通过留一交叉验证和五折交叉验证对该模型进行了验证，AUCs分别为0.9034和0.8954 ± 0.0030[15]。

骆嘉伟等人通过已知的微生物-疾病关联网络连接疾病相似性网络和微生物相似性网络，构建了一个异质性网络，然后通过整合网络拓扑相似性并基于随机游走算法来计算微生物-疾病关联得分，在该方法中骆嘉伟等人将每个微生物-疾病关联对看作一个不同的关系层次,并根据网络拓扑相似度赋予不同的权重，在对HMDAD数据集中数据采用该方法进行留一交叉验证和五折交叉验证时，平均AUCs分别为0.9070、0.8896+0.0038[16]。

吴桐等人借助美国公开肥胖人群数据库提取了肥胖人群的特征，此特征主要显示肥胖人群与正常人群的差异,他们借助属水平的肠道菌群数据建立了肥胖分类机器学习模型,并进行网络搜索确定最佳模型参数,并通过梯度提升回归树进行疾病预测,结果的AUC达到了0.725725[17] 。

## 研究内容及关键问题

### 1.3.1研究内容：

本研究通过使用机器学习的方法来寻找肠道微生物与疾病之间的关联关系，以根据这些关联关系判断用以采样肠道微生物的样本健康与否，即进行样本分类任务。图神经网络是本课题的创新内容，肠道微生物之间存在着相互作用关系，而由于菌群结构和丰度的区别，这样的关系在具体的每个样本中又有所不同。本研究为每个样本根据他的肠道微生物丰度测序结果构建微生物相互作用网络，以作为图神经网络的输入，使用图分类的方法来讲该图归为某一特定类别(疾病)，以完成疾病诊断任务。

除了疾病诊断以外，本研究的另一个核心任务是根据每个样本的具体情况给出个性化的诊疗建议，这部分将参考疾病诊断任务中分类器的表现，选择表现最佳的分类器来给出诊疗建议，这部分需要研究的问题是，如何获取足够有说服力的数据来源，因为本研究将微生物群落之间的关系作为输入，因此首先需要构建微生物关系网络，而且应该有足量的输入数据以供训练，验证和测试。

其次，如何构建模型，模型中的超参数，激活函数，模型中各层次的内容，各层次间的连接方式，以及使用图神经网络（Graph Neural Network）进行图分类任务时的图卷积算子和图池化算子的选择也是研究的关键内容。

最后，需要尝试各种优化模型效果的方法来改善分类器的效果，以使本研究有实用性价值。

### 1.3.2关键问题：

（1）构建图神经网络

拟定采用图分类的方式来构建图神经网络模型，模型工作流程中的卷积层，池化层，都需要进行测试并选择。分类器一般而言使用全连接神经网络作为分类器。

（2）模型训练数据来源

也就是微生物群落关系图的生成，目前生成关系图的被人们广泛认可和采用的方法为：首先计算微生物之间的相关度（例如：皮尔森相关系数和斯皮尔曼相关系数）并以两两微生物之间的相关度生成邻接矩阵，其次设定一个阈值，保留该邻接矩阵中大于该阈值的项。本部分打算借助近年来新提出的方法，来根据微生物组OTU表使用邻边预测与传统阈值计算结合的方法，自动生成一个邻接矩阵来作为模型的输入数据。

（3）预测准确率问题

现阶段有大量的机器学习模型可以应用于本任务，如何选择最适合本研究的模型是一个重要问题。模型的参数优化需要进行大量的实验，并参考先前进行过的实验；模型的泛化能力也是在实际应用场景中不得不考虑的因素，也需要使用恰当合适的评价指标来评判模型性能；在模型训练阶段可以使用dropout等方法来略微提升模型性能，在本文中也有所体现。

## 本文组织结构

本文共分为四章。其中第一章主要阐述了肠道微生物，疾病预测相关的研究进展，还有国内外机器学习方法进行研究的现状。第二章主要介绍算法部分是本论文的核心工作，主要讲解了如何将五种精心设计的分类器应用于六个用于疾病分类的数据集上，包括数据预处理，构建微生物相互作用网络，模型参数调整，结果分析和展示等。第三章处于实际应用的考虑，设计了逆因果算法，使得本系统可以在给出患者诊断结果的同时，给出微生物丰度水平的诊疗建议。第四章则介绍了系统的设计和实现方式，同时完整的测试了系统能够提供的所有功能模块，以确保能够投入使用。

**第2章 基于机器学习的肠道微生物与疾病关联关系分析方法**

## 2.1数据集来源与构建

### 2.1.1 数据来源

本研究共囊括2424个样本，这些样本源自于八项研究六种疾病，测序数据来自于鸟枪法宏基因组学。Cirrhosis数据集由118例患病样本和 114例健康样本组成; Colorectal数据集中，直肠癌患者共48例，数据集中健康样本与小直肠瘤作为对照样本，共73例; IBD数据集中，疾病样本由溃疡性结肠炎和克罗恩病组成，共25例，健康人共有85例; Obesiy数据集包含164例肥胖症病人和89例消瘦人士;T2D数据集包含了170例来自中国的二型糖尿病患者和174例健康人，WT2D数据集样本以类似的方式采样欧洲样本，共计有53例患者样本和43例对照样本。具体数据见表2.1.1。

表2 . 1 . 1 肠道微生物与疾病数据集

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 数据集 | 疾病类别 | 健康样本个  数 | 患病样本个  数 | 包含物种个  数 | 参考文献 |
| Cirrhosis | 肝硬化 | 114 | 118 | 542 | [18] |
| Colorectal | 结直肠癌 | 73 | 48 | 503 | [19] |
| IBD | 炎症性肠病 | 85 | 25 | 443 | [20] |
| Obesiy | 肥胖症 | 89 | 164 | 465 | [21] |
| T2D | 2 型糖尿病 | 174 | 170 | 572 | [22] |
| WT2D | 2 型糖尿病 | 43 | 53 | 381 | [23] |

### 2.1.2 数据预处理

原始的测序数据为碱基序列，无法直接用于丰度相关分析，所以需要将测序得到的碱基序列构建为微生物丰度数据。Pasolli 等人先前开展过类似的工作，使用上述数据集构成了一个微生物丰度矩阵，所以，首先的需要将该矩阵拆分为分别的六种疾病数据集，此后需要使数据集具有可计算性，例如剔除从未出现过的数据等。本研究中只用到了肠道微生物丰度数据，以及样本标识数据（年龄，性别以及国籍等），可以将分离出来的六个数据集中无关的信息抛去（如测序时间等）。以此法保留的数据为一个csv格式的操作聚类单元（OTU表），每一行是一种OTU（在经过比对后被认为具有相同的基因序列），经过标注后对应于真实世界中的某一种细菌，每一列为一个测序样本，即采样于某一个患者或者健康人的肠道。此外，为了提高训练效率，应该删除在所有样本中都为同一丰度的值，这样可以使得数据中重要的可分信息被保留下来，从而提高模型识别能力。此外，由于菌种的规模千差万别，某些菌的丰度可以极大，与其他菌的丰度差距可以达到十个数量级，这样的菌会极大程度上干扰模型的分类能力。因此，首先对每一种菌的丰度取对数避免过大的差距干扰识别特异性信息，其次要进行归一化操作，将菌的丰度区间缩小至0至1的区间内，避免差距悬殊的丰度影响模型性能。

此外，本研究中使用了图分类的方法来辅助诊断。也就是说，输入数据必须是一张图，图上的每个节点代表一种菌落，图中的边代表菌和菌的相关性，本研究计算两种菌之间的相关度得分，当得分大于某个阈值时两菌之间存在一条边，反之亦然。训练时每个样本都生成一个图，样本的健康或者患病标签就是图的标签。为此，需要设法将数据集中的微生物丰度矩阵转变成若干个模型能够识别的图。经过上述预处理操作后，剩余的微生物都被认为具有分类价值，应当作为图中的节点，图中的边因样本而异。为了最大程度使得相关度阈值具有评判相关性的能力，本研究使用了MENA方法[24]来计算相关度阈值，该方法首先设定一个步长，即对每个阈值在0到1之间计算皮尔森相关矩阵的特征值的分布，起初特征值分布为高斯分布。而某一阈值以后的特征值分布都是泊松分布，这一阈值被本研究作为默认相关度阈值。这种基于随机矩阵理论的阈值划分方法在其他领域已经有了较为成熟的应用，因此划分阈值的准确性较优秀。

为每个样本所对应的图构建边时还进行了进一步处理，现有的相关度计算方法需要大量的样本才能较好的评判两种菌之间的相关度，无法满足每个样本都个性化定制图的需求。而且不同人群的肠道微生物组的丰度差别极大，因此有必要让模型能够在健康人与病人上都有较好的预测能力，本研究中使用将健康人与病人划分为多个集合。

为了解决上述问题本研究在读取数据集时，将健康样本和患病样本分为两个矩阵，分别计算这两个矩阵中每种菌之间的相关度，从而得到健康相关度得分矩阵和患病样本相关度得分矩阵。将这两个矩阵相加，再使用事先选择好的相关度阈值筛选，从而得到一个值为0或者1的邻接矩阵，0代表无关，1代表相关。上述由整个数据集生成的图被称为极大图，针对某一个具体样本，首先选择该样本中丰度不为零的菌，随后在极大图中根据这些菌所对应的节点取子图。也就是说每个样本对应的图，都是极大图的一个子图。除了边来自极大图以外，子图节点特征为该节点所对应的菌在样本中的丰度。由于病人体内的菌群结构与健康人不同也就是说病人子图所对应的节点与健康人子图所对应的节点不同，健康人与病人的子图结构会有所不同，例如：沙门氏菌会导致某种疾病，虽然在极大图中沙门氏菌被预测与许多菌群相关，但是在健康人体内没有沙门氏菌，也就避免了健康人子图结构中包含病人子图结构的问题。这在一定程度上使得模型具备了区分病人结构与健康人结构的能力。

## 2.2研究方法

本章研究的内容是使用机器学习的方法来寻找肠道微生物与疾病之间的关联关系，以根据这些关联关系判断用以采样肠道微生物的样本健康与否，即进行样本分类任务，以为后续提供诊疗意见和在实际场景中辅助诊断打下基础。

### 2.2.1 机器学习方法

本研究首先尝试将传统的成熟的机器学习方法应用于该任务中。具体而言本研究应用了基于sklearn库的随机森林模型与支持向量机模型。这两种模型的参数设置受到了Pasolli等人的启发。

首先采用的方法是支持向量机(SVM)，SVM旨在找到一个能够最大化不同类中样本间距的超平面，以使得超平面两侧的支持向量到超平面的几何距离最短。虽然SVM本质上是找到一个线性决策边界来分类样本，但是可以使用核函数来将样本扩展到更高维空间以找到用来划分样本类别的非线性决策边界。本研究中输入数据为一个样本的微生物丰度测序结果，共有n个输入向量，且他们的形状形状都是m\*1，其中m是微生物数量，n是样本数量，由于微生物丰度千差万别，很难以做到线性分类。因此在本研究中，采用径向基函数(RBF)来将样本特征向量映射到高维空间以寻找非线性决策边界，而RBF中的γ决定了每个向量的影响范围，当γ取值过小时，每个向量的影响范围都特别大，导致决策边界趋于线性，反观γ过大时每个向量的影响范围过小，模型容易过拟合，因此γ是一个需要在实际研究中不断调整的超参数。除此之外，SVM中由软间隔非线性分类模型产生的规则化参数C也直接影响着模型表现，C反映了软间隔分类时放弃离群点的程度，当C趋于无穷大时不允许出现任何误分类于是又变成了硬间隔分类；反观C趋于0时，不再关注分类是否正确，离群点的数量为任意，只要支持向量到决策边界的间隔尽可能大即可，因此也是决定SVM性能至关重要的参数。本研究中SVM模型的参数C和γ被分别从这两个集合中选择{2−3, 2−1, …, 215} 和 {2−15, 2−13, …, 23}，并据此训练SVM模型。随后基于五折交叉验证评判模型分类效果好坏，在五折交叉验证中样本被随机划分为了5个等大小的子集，在五折训练中每一份样本轮流作为验证集，来衡量在剩下的四份样本上训练出来的模型的分类能力。由于在某些数据集上正样本较少（IBD数据集中正样本比例为0.23），本研究特别地使用了分层交叉验证以保留原始数据集中每类样本的百分比。随后，一个单独子集被用来训练模型，其余子集被用于训练模型，这个过程重复了5次。最终测试结果取五折验证的平均结果。在评估SVM的最终结果时，在每个数据集上都以这样的方式尝试多次，选择acc最高的结果作为SVM模型s在该数据集上的最终结果。

本研究中还应用了一种名为随机森林算法（RFs）的集成学习算法，这种算法在训练时构建很多决策树，算法的最终是综合考虑各个单独决策树的结果。在本研究中树的个数为500，构建指标为基尼不纯度，它是被用以考量被错误分类的样本占总样本比例的指标，训练时不限制树的深度，直到所有样本都被分类为止。训练每一棵树时样本被有放回的抽样也就是训练时不采用所有样本而是随机抽取一些样本训练，同时对于每一刻决策树，他只选取所有特征的一个子集来训练，而且不同决策树选择的子集可能不同，这在一定程度上避免了对数据的过拟合问题。在测试时，通过平均每个决策树的预测来进行预测。构建决策树时会尽量使基尼不纯度最低，训练的停止条件是训练集中的所有样本都被划分出一颗独立的决策树，换言之如果仍有样本未被划分，那么深度可以无限延伸，这样会使得每一棵决策树对数据集拟合程度较高，但是随机森林算法会取各个决策树的平均结果，从而避免了某一棵树对训练数据的过拟合。

实际上，数据集中有些菌的丰度很低，且只在个例样本中出现，可能会影响模型训练效果。恰巧随机森林算法中内部嵌入有特征重要度计算算法，本研究利用这些重要度值来执行特征选择策略。具体流程如下：

1. 将随机森林算法应用于数据集中所有可被采纳的特征上

2. 依据计算出的重要度值将特征排序

3. 随机森林算法被在前k重要的特征上重新训练，k取值于如下集合{5，10，20，30，40，50，60，70，80，90，100，125，150，175，200}。

在评估随机森林算法的最终结果时，在每个数据集上都以这样的方式选择特征后以五折交叉验证的方式来评估模型的分类能力，选择acc最高的结果作为随机森林算法在该数据集上的最终结果。

### 2.2.2 深度学习方法

神经网络是模仿人脑中有着复杂相互作用关系的生物神经网络而构建的一种计算模型。神经网络中的每个节点(神经元)在接受信息，处理信息和传递信息的同时受到训练，可以对接收到的某种特征信息产生特定的反映。一个典型的神经网络由三个部分组成，分别是输入层、隐藏层和输出层。 全连接神经网络中，每个节点都与上下两层的所有节点相链接，每一层由若干个节点(神经元)组成，并且每个节点都有自己的权重系数，输入该层的数据在乘以权重系数后加上偏移量作为下一层的输入。从模型的三个组成部分角度分析，输入层节点数量一般与输入数据的特征维度相同，负责接收数据并传递数据。隐藏层可以有多个，隐藏层的职能是将数据升维以使数据线性可分，它负责处理输入层传来的信息。而输出层则负责将隐藏层传来的特征转化为直观的可被理解的信息。例如在分类任务中输出层的每个节点常被定义为某一类。介绍至此，神经网络中的每个隐藏层都对输入数据进行线性变换，这样只能处理线性可分的数据，而不能处理线性不可分的数据，为此需要引入非线性的激活函数以使得网络能够处理线性不可分数据，如Sigmoid，Relu激活函数等。全连接神经网络示意图如图2-1所示：

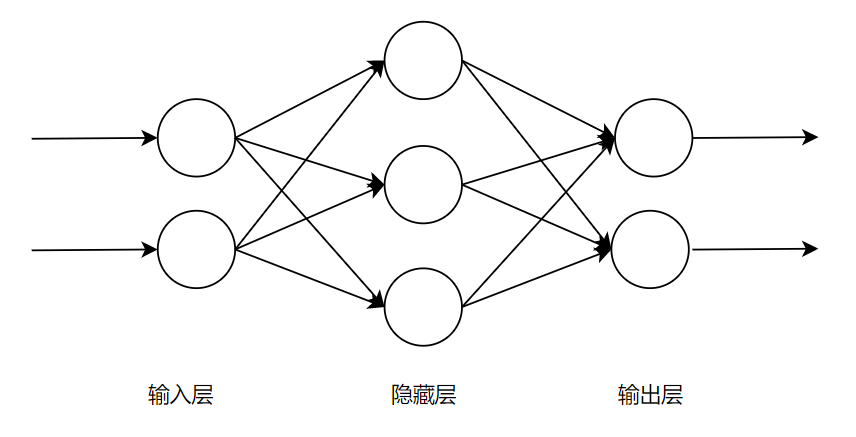


图2-1全连接神经网络示意图

全连接神经网络的一个核心算法是反向传播，正向传播的过程中，数据通过一层层隐藏层和激活函数，形成了被规定具有实际意义的输出向量，并基于该向量得出一个预测结果。在反向传播的过程中，神经网络将该预测结果与真实的数据结果进行对比，得到模型预测的误差，并据此反向遍历网络中的每一层求得其中权重关于损失函数的梯度，并根据学习率进行梯度下降以更新每一个神经元上的权重系数，目的是最小化损失函数使得模型具有分类能力。

卷积神经网络与全连接神经网络都由若干层节点组成，不同于全连接神经网络，卷积神经网络的每一层只与上一层中的部分节点链接。卷积神经网络通常由卷积层和池化层组成。具体而言，每一个卷积层通过卷积核来将该层中的节点与该节点的相邻节点的信息聚合到一起以作为下一层的输入信息。池化层则是为了缩小特征向量的维度，以避免大量参数导致的计算开销，还能在一定程度上缓和过拟合。

以上两种深度学习方法在分类领域常被使用并且能够取得较好的效果，本研究基于tensorflow的keras框架设计了全连接神经网络和一维卷积神经网络模型，

具体构建方法如下。

全连接神经网络的构建中，选择使用双层隐藏层，形状为（m \* n）的样本矩阵（m，n分别代表微生物数量和样本数量）在预处理后输入模型，经过两个隐藏层抵达输出层，每个隐藏层都以relu函数为激活函数，在本研究中并未采用dropout，因为预测准确率反而会因为dropout的引入而有所下降，推测原因在于数据经历过预处理后，保留下来的菌都是对分类有贡献的菌，dropout会导致样本特征信息丢失，从而适得其反，此外，隐藏层节点数量随着数据集而有所变化，例如在Cirrhosis数据集上为512-256，模型示意图如图 2.2.2(a)所示。模型的输出被设置为了两个节点，即健康样本节点与患病样本节点。

卷积神经网络常被用在图像处理领域，他可以捕捉到每个像素之间的连续性信息，而本研究中，不同样本和不同样本之间的菌群丰度和结构可以有很大差异，尤其是健康人和病人之间的微生物丰度可能截然不同，因此传统的卷积神经网络并不适合本研究。研究表明，人体内肠道菌群会达到某种平衡[8]，达到平衡之后菌群的丰度水平趋于稳定，又由于数据集中所有样本体内的菌种丰度都是按特定顺序排列的，以微生物名为自变量，微生物丰度为因变量构建的函数理应有着类似的变化趋势，本研究应用一维卷积神经网络正是为了捕捉这种平衡下的菌种相对丰度变化情况。具体而言，本研究应用一维卷积神经网络处理数据集中的每一个样本，一维卷积核的大小被设置为3，填充方式为valid填充，即卷积核只在输入向量内部卷积，卷积步长为1，采用64个一维卷积核在形状为（m，1）的样本上卷积。在一维化之后输入128个节点的全连接层中，最后输出层一样为分别代表健康，患病的两个节点，具体模型如图2-2所示。

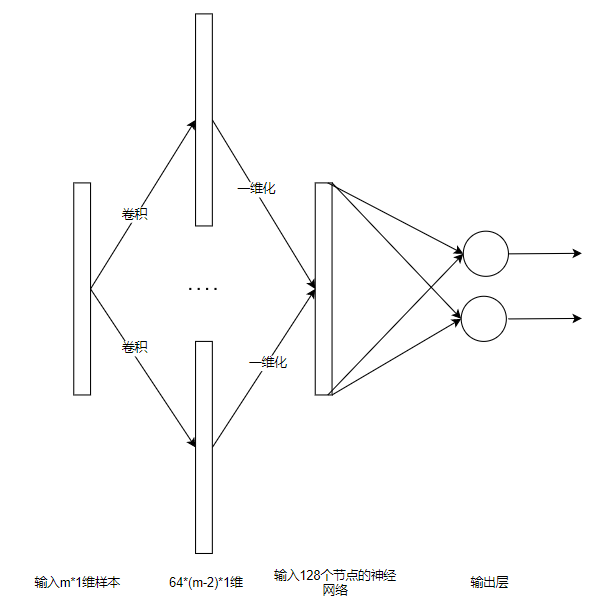


图2-2 一维卷积神经网络示意图

以上两种神经网络的输出使用softmax函数转换为可供损失函数分类的向量，损失函数选择分类任务常用的交叉熵损失函数，优化器为可以根据训练轮数自动调整学习率的adam优化器，以accuracy和ROC曲线下面积（AUC）作为评价指标。其他参数如batchsize为16，epoch一般为5轮，学习率使用默认设置1e-3，此时的loss已经小于0.1，认为模型足够收敛，并且能够取得不错的准确率。

另一个问题是，本研究中每个数据集的样本量都介于100到400，当运行次数少时，在较大程度上运行结果会受到数据的影响，为了能够准确地评估模型的真实预测能力，本研究应用了分层10折交叉验证，取平均值作为最终运行结果。同时对于每一种模型，以相同的比例划分训练集和验证集。以得到模型之间相对精准的对比结果。

以上提到的所有方法都是直接对微生物丰度表进行处理，试图从微生物丰度中提取特征信息，以判别样本是否患病。虽然上述方法最终都取得了不错的表现，也说明微生物丰度中的确包含了关于样本健康情况的信息，但是却过分局限与某些微生物的具体数值，以至于忽视了肠道菌群在相互作用下达到的动态平衡。“人类微生物组计划”研究表明，成人身体健康没有罹患疾病时，占据主导地位的肠道菌群为厚壁菌门，同时这些菌的丰度较为稳定。反观身体亚健康或生病时，厚壁菌门丰度会有较大波动与平时不同。孕妇产后体内的放线菌门、变形菌门的菌种丰度较同龄人有大幅度增加[8]，产后精神萎靡、产后抑郁等症状可能与肠道菌群丰度波动有一定的关联关系。由此可见，人体内的微生物群落间的相对丰度已经达到了某种平衡，这种动态平衡被打破时，人体很有可能患疾病。而这种动态平衡很难由某几次测序的丰度表示，更重要的是组成这种平衡的菌群结构及其相互作用网络。因此，本研究同时使用肠道微生物丰度和肠道菌群结构信息来预测微生物疾病之间的关联关系，再根据这些关联关系进一步判断用以采样肠道微生物的样本健康与否。

传统的机器学习方法桎梏于处理欧式数据，也就是输入具有固定形状的数据，例如输入全连接神经网络的数据需要有相同的特征维度。健康人和患病人体内的肠道微生物丰度和结构信息有着较大的区别，例如在肝硬化病人中，与肝硬化疾病关联最大的韦荣球菌、链球菌等菌类丰度较高，而健康人体内这两种菌的丰度几乎为零。这就导致每个样本体内的菌群结构都有所不同，这样的数据被成为非欧式数据，而本研究应用的图神经网络恰恰可以解决这一问题。

图神经网络的输入数据一般是一张图，图中可以有任意多个节点记录节点特征，每个节点都对应着非欧式数据的某一特征，而图中的边则代表了节点之间的关系，存在边说明节点之间存在着某种关系，在图神经网络学习信息的过程中，节点特征会沿着边不断进行信息传递和节点特征更新，这样以来图神经网络不仅能够学习输入数据的数值信息，更能学习输入数据的结构信息。本研究搭建图神经网络以图分类的方法来研究某个样本的微生物丰度信息以及该样本的微生物结构信息。具体而言，本研究中的图神经网络的输入数据为某一个样本的微生物菌群结构图，该图中的节点为具体的某一种微生物，节点之间的边为微生物与微生物之间的相互作用关系。在经过图神经网络的学习后，将整个图上的信息整合为一个节点，并将该节点用作分类任务的输入，即进行图分类以分辨输入神经网络的样本患病与否。

为了使微生物菌群结构图中的边具有实际意义，本文使用相关度计算的方法来计算两种菌之间的相关度得分，得分大于某一阈值时视为这两种菌相关，并在图中为这两种菌添加一条边。而依据相关度构建有时过于绝对，尤其是临近相关度阈值的一对菌的关系难以界定。因此，本研究用MENA方法来确定相关度阈值，如2.1.2数据预处理中所述，该方法寻找使得相关矩阵的特征值的分布发生改变的阈值，以该阈值为衡量一对菌相关与否的阈值，这一方法解决了阈值划分过于绝对的问题。而计算相关度的方法，本研究采用了皮尔森系数(Pearson correlation coefficient)，斯皮尔曼系数(Spearman’s correlation coefficient)[25]以及最大互信息系数（maximal information coefficient）。皮尔森系数与斯皮尔曼系数是两种经典的相关性计算方法，在经过游懿君等人的实验后[26]，较Sparcc与CCLasso等微生物领域的常用相关性计算方法能够更好的发现微生物相关度，而最大互信息系数MIC可以捕捉到复杂的非线性关系，它可以很好的作为两种传统方法的补充,最终构建的网络如图2-3所示，其中图(a)为在pvalue为0.001情况下构建的网络，图(a)为在pvalue为0.01时构建的网络，为了使得信息损失不过于严重，本研究采用pvalue为0.01时的阈值为最终阈值。

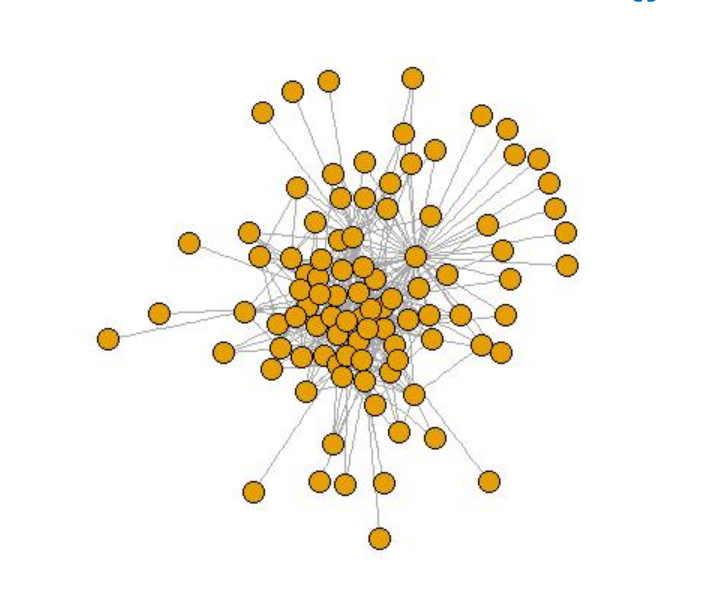
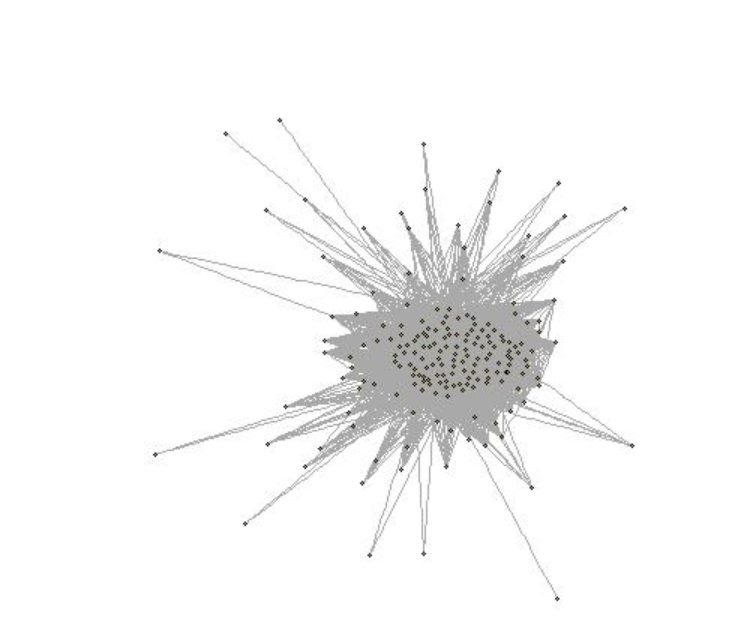
 

图2-3(a)MENA构建的网络(p=0.001) 图2-3(b) MENA构建的网络(p=0.01)

除了输入输出以外，图神经网络的模型通常还有图卷积层，图池化层以及分类器三个重要部分。图卷积层如同卷积神经网络的卷积层，每个节点聚合周围邻居的信息并将该信息整合自身信息后封装为自身新的节点信息，该信息被成为节点嵌入，图卷积的过程就是不断进行信息聚合和消息传递的过程，类似于卷积神经网络中采用等值填充后的效果，图卷积时只会不断地更新节点嵌入以尽可能地使节点嵌入包含整个图的信息，而不会改变图的结构。而最终分类时我们希望无论是怎样的图最终都能够被表示为欧式数据以供分类器学习，因此需要引入池化层。图池化层是将节点不断聚簇使其坍化为一个代表了整个图的节点的过程，最终该节点的节点嵌入被称为图嵌入。最终的分类层则比较简单，通常使用全连接神经网络来处理图嵌入，将其归为某一特定类别。本研究中对图卷积和图池化方法做了较多的尝试，本研究首先针对现有基准数据集搭建网络，使用不同的数据集分别衡量图卷积层，图池化层的效果，在基准数据集上取得较好的结果后才尝试运行自己的数据集。

对于图卷积算子，首先尝试使用图卷据领域最经典的Graph Convolutional Networks[28] (GCN)模型完成卷积任务。本研究基于PyTorch Geometric框架实现了GCN模型，在Cora节点分类数据集上测试了使用完成卷积后的节点分类，accuracy为78%。但是在后续研究中发现GCN是基于谱域的方法，也就是说它在信息聚合过程中关注整个图的邻接矩阵（拉普拉斯矩阵），这意味着每当加入新节点后，都需要所有节点参与拉普拉斯变换以重新计算图的特征向量，这不但使得训练模型时计算量大而且导致使用GCN搭建的模型不具备处理训练时没有见过的节点的能力。而我们的任务中由于人体肠道菌群结构差距很大，极有可能出现新的菌群（作为新的节点），所以GCN并不适用于我们的任务。

为了解决以上问题，尝试了使用GraphSAGE[29]作为卷积算子，它是一种基于空域的归纳学习算法，基于空域是指训练时仅关注每个节点的邻居节点，归纳学习是指可以对训练过程中见不到的数据直接计算而不需要重新对整个图进行学习。总的来说，GraphSAGE学习的内容不再是为每个节点定制嵌入生成策略，而是学习泛用于所有节点的信息聚合函数，图中的每个节点都可以用这个聚合函数来整理信息和传递信息，甚至从未出现过的节点都可以使用该聚合函数。因此GraphSAGE的泛化性好，对新加入的节点，可以直接套用训练好的信息聚合函数，而不必对重新更新整个图的节点消息传递策略，同时通过堆叠聚合函数的层数，可以使模型学习到多跳邻居的信息。GraphSAGE中某一个节点A更新自身嵌入的流程如下：

1.令l∈[1,2,…,L]，L指的是采样深度，也就是最大邻居跳数，L=1时只采样直接相邻的节点，L=2时也采样邻居的邻居，以此类推。对于节点A的第l跳邻居，随机采样不多于Sl个l跳邻居节点，Sl代表了第l跳邻居的最大数量，也就是说处于效率考虑GraphSAGE并不采样全部邻居节点信息。

2.从l-1跳邻居开始，l-1跳邻居聚合l跳邻居节点的嵌入以生成自身节点嵌入；接下来l-2跳邻居聚合l-1跳邻居的节点嵌入，以此类推直到节点A聚合到了自己直接相邻的节点的嵌入。

3.节点A依据自身节点嵌入与聚合到的邻居节点嵌入计算节点A新的节点嵌入。

上述流程中涉及到两个重要的操作就是邻居节点聚合和节点嵌入计算。

首先讨论邻居节点聚合，为了完成这个操作，需要有一个聚合函数来帮助节点A整合邻居信息，在图中每一个节点的邻居是无序的，所以能够有一个对称的聚合函数（即也就是对它输入的各种排列，函数的输出结果不变）以确保被训练的聚合函数可以被应用于任意序列组合的邻居节点特征集合上。本研究采用的聚合方式为均值聚合，即取上一层邻居节点嵌入的平均值。可以看出GraphSAGE的信息聚合是分层进行的，每一层对应了顶点的每一跳邻居。

节点嵌入计算方式为将聚合得来的邻居节点的嵌入乘以聚合权重矩阵的结果与自身节点乘以自身节点权重矩阵的结果相加，再输入非线性激活函数。具体节点嵌入的计算方式如公式2-1所示。

2-1

公式中：为当前层数第层可以看作是在计算倒数第跳邻居，为非线性激活函数，为上一层邻居节点嵌入的平均值(为节点的邻居节点的集合)，是节点在上一层的节点嵌入，即为当前待生成的新的节点嵌入。其中和是可以用来学习的参数，分别是邻居节点的权重矩阵和节点自身的权重矩阵。

在本研究中，GraphSAGE模型中的每个节点都聚合两跳邻居，以避免邻接矩阵过于系数而无法充分学习信息。GraphSAGE在Cora上的表现与GCN十分接近，优势在于它可以处理新的菌群，而且当新的节点出现时不需要如同GCN一样将现有结果推倒重来，而是可以直接应用已被训练过的信息聚合函数。

考虑到不同菌群的规模差距较大，从而图结构中节点特征差距较大，根据微生物相关度计算方法，丰度较大的节点更有可能被计算为与其他节点相关，于是这部分节点可能被过分学习，导致丰度较小的节点信息不能很好的被模型学习到，然而在实际场景中，不同的邻居节点可能对核心节点起着不同的作用。为了解决这个问题，本研究引入注意力机制，使用Graph Attentlon Networks[30]来作为图卷积算子，GAT就是带注意力的对节点进行聚合，这样一来模型可以学习到对不同节点的权的分配（即注意力），从而提高了模型的准确率。GAT也同样可以被描述为信息聚合和计算节点嵌入两部分。信息聚合时，GAT简单的采样所有邻居节点的特征，靠着注意力系数来带权重地整合邻居信息。GAT地重点在于节点嵌入生成方式，同样的以一次信息聚合和消息传递为一层，GAT在已有当前层地节点嵌入hi试图求下一层节点嵌入的方法如下：

首先计算节点i的所有邻居节点(j∈Ni)和节点i自身的相似系数，计算方式如公式2-2所示：

2-2

公式中W是一个在个节点之间共享的权重矩阵，W∈R，他做了一个线性映射以对节点特征增维，则对于顶点i，j变换后的特征进行了拼接，a则把拼接后的高维特征映射为一个实数，显然GAT通过可学习的参数W和a来计算结点间的注意力。

其次通过相似系数计算注意力系数，两个节点i，j之间真正的注意力系数其实就是相似系数的归一化，通俗的理解就是softmax化相似系数，同时做了一些非线性变换，计算方式如公式2-3所示

2-3

其中是节点i的邻居节点，LeakyReLU是一个非线性激活函数是ReLU激活函数的变体，exp代表取以e为底的指数。

最终GAT网络还采用了多头注意力机制来根据得到的注意力机制带权的计算每一头产生的节点嵌入，最终将这些节点嵌入连接之后取平均得到最终注意力。这使得节点注意力的学习更为平均，使得注意力机制被较好的应用到图神经网络中。具体实现方式如下：K为总注意力头数，k为当前第k头注意力，为当前层节点j的节点嵌入，为即将得到的节点i的节点嵌入。代表取连接并取平均值，代表对i的所有邻居顶点j，计算结点i，j之间的的第k头注意力系数，与上述公式一样，为对第k头注意力而言需要学习的权重矩阵，,in\_dim为输入该卷积层时节点特征的维度，out\_dim为卷积完成后节点特征的维度，具体如公式2-4所示

2-4

很明显的，在GAT中输入的样本数量为1，也就是说，GAT网络每次只处理一个图，他是无法直接按批处理数据的。在本研究中，输入样本的形状为m\*n，其中m为微生物数量，n为样本数量，在实际训练时，每个样本都生成了一张图，于是每次训练时每个节点代表一个微生物，丰度自然也是一维，传入卷积层的节点特征形状都是(1\*1)，前者代表该batch中只有一张图，后者代表节点特征为1，可以看到在图神经网络中由于节点特征较少，可以更充分的学习到图的结构信息，这与我们采用图神经网络的预期相符。

上文中的GraphSAGE有着自己定义的按批计算的矩阵化实现，但是为了方便更换卷积算子而不用重新搭建模型。本研究中GraphSAGE与GAT一样也一次输入一个图，但是本研究仍然实现了batch化的训练。具体而言，假定batchsize为16，那么就串行的训练16张图，每次训练计算loss之后不进行反向传播，而是将loss累加，直到batch结束之后统一进行一次反向传播。实际测试表明，不进行batch化训练时，loss和acc需要训练数百轮之后才开始改变，而验证集上的acc几乎不变，这说明按batch进行训练是很有必要的。

GAT虽然在训练多头注意力时需要很大的计算量，但是GAT自带有并行在多个节点上训练的能力，从结果上看，GAT比GraphSAGE的训练时间更长，约为后者的三倍。测试时，在Cora数据集上GAT取得了最好的准确率（84%），这说明GAT在数据集上有着最好的卷积能力。此外，GAT 是基于空域的方法，也就是说它只关注邻居节点，而不需要关注整个图结构。

上述操作都属于图卷积操作，是为了整合整个图的信息道具体某节点，而本研究进行的是图分类工作，因此需要获得整个图的表征方式，也就是图池化。图池化操作是将相似的节点簇聚合为单个节点，以降低图嵌入表示的维度的过程，这个过程是为了将图上每个节点嵌入的信息整合起来以得到整个图的嵌入表示。在选定卷积算子后，本研究中首先尝试了对所有节点嵌入简单取平均值来作为分类任务的输入，并在图分类数据集上进行测试来记录效果。结果显示GAT与这种简单的全局平均池化只能在6分类蛋白质结构数据集ENZYMES上取得30%的精确率，结果较差。为了使得图结构信息以及节点嵌入信息能够更好的保留，本研究采用了DiffPool[31]方法，这是一种层次化的图嵌入生成方法，DiffPool是一个单独的图池化层，他被安排在两到三个图卷积层之后，它将经过图卷积的邻接矩阵与DiffPool层中以图卷积方式学习到的分配矩阵相乘以得到被粗化的邻接矩阵，也就是将相似节点通过分配矩阵聚成一簇，并对被粗化过的图进行卷积求每一簇新生成节点的嵌入，如公式2-5所示。

2-5

其中为当前层，为邻接矩阵，为节点嵌入，重复池化过程一到两轮，此时整个图被层次化的池化为了一个用于分类任务的节点，该节点的特征就是图嵌入。所以DiffPool最为关键的任务就是如何找到每个节点在下一城中属于那一簇。DiffPool同样以图卷积的方式学习一个生成分配矩阵，如公式2-6所示。

2-6

其中为当前层节点数量，GNN为图卷积算子，在本研究中为GraphSAGE。于是某一个Diffpool层中新的邻接矩阵和节点嵌入的生成方式为(为求矩阵乘积)，如公式2-7和2-8所示。

2-7

2-8

可以看出，Diffpool使得相似节点被有规则地聚簇，为确保每个簇中的节点是确定的,该模型还对学习得到的分配矩阵进行限制,使得分配矩阵地每一行尽可能地是one-hot向量(每行只有一个1,其他都是 0 的向量)。本池化模型在基准数据集上取得了较好地效果，具体而言，在PROTEINS数据集上准确率达到了76%，在ENZYMES上准确率可达58%，该模型的缺点是模型结构复杂，收敛慢，因此需要大量的训练才能使得模型结果较好。优点在于该模型可以自由地与图卷积层搭配，从而使得模型效率发挥到最佳。

## 2.3 结果分析：

上一节详细介绍了五种模型的设计思路，而在实际评判模型性能时还采用了一些优化方案来提高模型分类能力。本节将着重评估模型的分类能力以及采用的优化方案对模型效果的影响。首先，本节在六个数据集上测试了5种模型，分别是随机森林(RF)，支持向量机(SVM)，图神经网络(GNN)，深度神经网络(DNN)，一维卷积神经网络(1D-CNN)，其中图神经网络由于模型复杂训练时间过长，采用了五折交叉验证，其余四种模型使用更为严苛的十折交叉验证。

模型参数设置如上一节所述，对图神经网络网络而言，上一节分析中GAT和GraphSAGE作为卷积算子在基准数据集上都取得了较好的结果，然而本研究最终使用的卷积算子是GraphSAGE，GAT虽然在基准数据集上取得了较好的效果，然而在实际应用到本任务时却表现不佳。以Cirrhosis数据集为例，在五折交叉验证中下，分类器预测的auc达到0.9以上而acc只有0.6左右而且训练时loss只能下降到一定程度就不再继续下降，训练时的acc也上升至0.8边持续波动不再上升，auc却可以稳定在0.85以上，原因可能在于每个样本生成的邻接矩阵过于稀疏，在171396个可能出现的边中，包括对角线元素的自旋只有7500个被预测出来的边，所以GAT能够学习到的信息十分有限，虽然能够较好的处理学习到的信息，但是受到数据本身的桎梏很难学习到足够的信息。总而言之，由于自身算法设计原因GAT只能处理一跳邻居，而GraphSAGE可以处理多跳邻居，在一定程度上解决了数据集从丰度矩阵转化到图时的信息损失，因此，GraphSAGE最为适合本研究，在接下来的测试中也一律选择GraphSAGE作为图卷积算子。模型测试结果如表2-1所示。

表2-1 五种模型评估结果对比

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Cirrhosis | Colorectal | IBD | Obesiy | T2D | WT2D |
| RF | accuracy | 0.877 | 0.805 | 0.809 | 0.644 | 0.664 | 0.703 |
| auc | 0.945 | 0.873 | 0.890 | 0.655 | 0.744 | 0.762 |
| SVM | accuracy  auc | 0.792  0.884 | 0.644  0.612 | 0.773  0.830 | 0.640  0.593 | 0.651  0.663 | 0.606  0.583 |
| GNN | accuracy | 0.826 | 0.644 | **0.818** | 0.612 | 0.543 | 0.632 |
| auc | 0.877 | 0.642 | 0.875 | 0.659 | 0.567 | 0.764 |
| DNN | accuracy | 0.847 | 0.769 | **0.821** | **0.645** | 0.656 | 0.657 |
| auc | 0.936 | 0.815 | 0.874 | 0.647 | 0.705 | 0.710 |
| 1D-CNN | accuracy | 0.837 | 0.772 | **0.826** | 0.629 | 0.638 | 0.676 |
| auc | 0.922 | 0.817 | 0.890 | 0.635 | 0.688 | 0.738 |

从模型效果上看，随机森林模型是当前表现最佳的模型，不仅在所有数据集上都取得了最佳auc，在四个数据集上也取得了最佳acc，这说明随机森林算法与本研究较为契合。支持向量机在所有疾病数据集上的结果都比如随机森林模型，只在部分数据集上表现得与随机森林模型接近，相对而言不适合后续在这些数据集上采用。深度学习模型在大部分数据集上的表现都略微逊色于随机森林模型，而在IBD数据集上的acc表现普遍优于随机森林模型，DNN模型在Obesity上也取得了高于随机森林模型的acc，而图神经网络仅仅在两个数据集上表现较好，在Colorectal数据集上的表现明显逊色于其他方法，原因可能在于该数据集中原本共有三种类别的样本，分别是癌症，肿瘤和健康。与Pasolli等人的研究一致，本研究中将癌症定义为阳性，其余两种定义为阴性。推测图神经网络表现不佳的原因如下，这两种被定义为阴性的人群也有着不同的肠道菌群结构，反而癌症和肿瘤样本的肠道菌群结构较为相似，这对于图神经网络而言是十分严重的干扰因素。

从实用性角度出发，在五种模型中，SVM模型训练非常快，可以在短时间内给用户响应，但是效果并不理想。其次快的是经典的随机森林算法，该算法训练效果较好可以在多个数据集上取得最高的acc和auc。但是在特定数据集例如IBD上，神经网络的表现明显更好。而图神经网络虽然可以在IBD数据集上取得超越随机森林算法的效果，但是由于模型复杂，且需要大量的训练才能收敛，并不适合作为实时运行的模型。因此图神经网络需要提前训练好效果最好的模型保存本地，在模型推理时，直接调用现成的训练好的模型。

本研究还分析了表现最好的随机森林模型的优越之处，以试图将其应用于其他模型以改善分类能力。随机森林模型在经过了特征选择的数据集上训练，并且表现超过了不进行特征选择的结果，结合对数据的观察发现，在数据集中某些微生物出现频率较低，只在个例样本中出现，与疾病之间并没有必然的联系，甚至可以成为干扰特征在参与训练的时候降低模型的分类能力，因此为了选取能够更加显著地影响模型分类能力的微生物，本研究降低参与训练的微生物数量，旨在使模型仅关注关键微生物，以尝试高模型的acc与auc。具体流程图如图2-4所示。

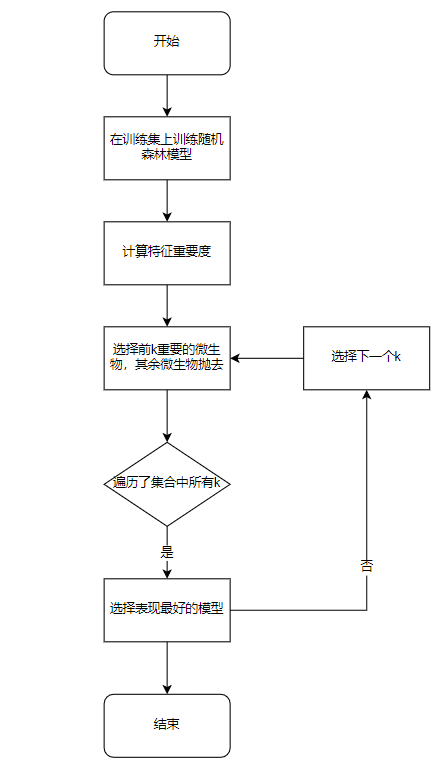


图2-4选择关键微生物训练流程

首先使用在整个数据集上训练好的随机森林分类器来获取微生物重要度，筛选前k重要的微生物保存下来，以供后续分类器真正的训练，k的取值在集合{5，10，20，30，40，50，60，70，80，90，100，125，150，175，200}中，每个数据集都选择在十折交叉验证中能够得到最高acc的k值来作为最终结果。除此之外，其余设置于前述设置保持一致，最终模型取得的结果如表2-2所示。

表 2-2 特征选择后的结果对比

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Cirrhosis | Colorectal | IBD | Obesiy | T2D | WT2D |
| RF | accuracy | 0.877 | 0.805 | 0.809 | 0.644 | 0.664 | 0.703 |
| auc | 0.945 | 0.873 | 0.890 | 0.655 | 0.744 | 0.762 |
| SVM | accuracy  auc | 0.799  0.862 | 0.636  0.645 | 0.786  0.841 | 0.645  0.633 | 0.612  0.639 | 0.599  0.589 |
| GNN | accuracy | 0.831 | 0.675 | 0.815 | 0.634 | 0.593 | 0.612 |
| auc | 0.879 | 0.701 | 0.860 | 0.657 | 0.587 | 0.744 |
| DNN | accuracy | 0.837 | 0.786 | **0.837** | 0.641 | 0.645 | 0.65 |
| auc | 0.922 | 0.802 | 0.893 | 0.643 | 0.712 | 0.695 |
| 1D-CNN | accuracy | 0.853 | 0.747 | **0.833** | **0.654** | 0.641 | 0.635 |
| auc | 0.931 | 0.799 | 0.852 | 0.616 | 0.707 | 0.710 |

表中结果显示，本方法可以提升某些神经网络的效果，全连接神经网络和一维卷积神经网络在IBD数据集上的表现在经过特征选择后超过了随机森林算法。在Obesiy数据集上仅采用前30重要的微生物进行训练时，一维卷积神经网络的表现在一定程度上甚至优于随机森林模型，虽然auc相对于不进行特征筛选时有所降低，但是acc超过了随机森林模型0.1。据此可以推测在Obesiy数据集上，对疾病起到关键作用的微生物可能只有30种。在仔细观察后发现，某些数据集上，该方法会降低模型的分类能力，例如在colorectal数据集上有三种分类器的表现都不如不进行特征选择，这说明在这些数据集上，微生物对疾病的贡献比较均匀，仅凭前k种微生物难以较好的找到微生物和疾病的关联关系，从而难以正确的诊断疾病。

最后，本节汇总了五种分类器在六个疾病数据集上能取得的最佳结果如表2-3所示，肝硬化数据集cirrhosis被模型最好的学习，在此数据集上地accuracy为0.877，auc为0.945；其次在IBD数据集和colorectal数据集上上表现较好，分别可以取得0.826的acc，0.89的auc以及0.805的acc，0.873的auc；相对而言，其他三种数据集并没有被很好的学习，acc的值介于0.6到0.8。从分类效果角度出发，随机森林算法在四个数据集上取得了最高acc与auc，是最为契合本研究的分类器，而且随机森林算法在所有疾病上都取得了最佳auc，一维卷积神经网络的效果在IBD数据集上可以超过全连接神经网络，原因在于本研究中微生物排序方式是有生物学依据的，所以卷积神经网络所捕捉到的信息是真实存在的微生物之间的联系。在IBD数据集上，特征选择的方法改善了模型训练效果，使得accuracy提升到了0.843而auc只降低了0.03，所以优化方法可以在一定程度上改善模型效果。

表 2-3 五种模型取得的最佳结果

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Cirrhosis | Colorectal | IBD | Obesiy | T2D | WT2D |
| RF | accuracy | 0.877 | 0.805 | 0.809 | 0.644 | 0.664 | 0.703 |
| auc | 0.945 | 0.873 | 0.890 | 0.655 | 0.744 | 0.762 |
| SVM | accuracy  auc | 0.799  0.862 | 0.644  0.612 | 0.786  0.841 | 0.640  0.593 | 0.651  0.663 | 0.606  0.583 |
| GNN | accuracy | 0.831 | 0.675 | **0.818** | 0.634 | 0.593 | 0.632 |
| auc | 0.879 | 0.701 | 0.875 | 0.657 | 0.587 | 0.764 |
| DNN | accuracy | 0.847 | 0.786 | **0.837** | **0.645** | 0.656 | 0.657 |
| auc | 0.936 | 0.802 | 0.893 | 0.647 | 0.705 | 0.710 |
| 1D-CNN | accuracy | 0.853 | 0.772 | **0.833** | **0.654** | 0.641 | 0.676 |
| auc | 0.931 | 0.799 | 0.852 | 0.616 | 0.707 | 0.738 |

总的来说当前机器学习方法十分成熟，本研究中机器学习方法仍然是泛用性最高的方法，深度学习方法乃至图神经网络方法在肠道微生物数据集上提升的效果不尽明显。仔细分析后，原因有三：

其一在于，数据集中的样本并不一定是完全合理的正确的样本，这些样本原来是rna测序数据，在经过基因比对后被标注为某种微生物的相对丰度数据，这些数据可能会出现假阳性和假阴性，从而产生噪声干扰模型训练，这个问题应该结合生物信息学的技术，采用更加先进严格的rna-seq技术来处理。

其二在于，数据集中的样本量较少，相比于机器学习，深度学习可以吸纳海量样本中的特征从而使得分类性能随着有效数据增加而不断加深，尤其是结构复杂的神经网络（如本研究中的图神经网络）。本研究中数据量较少不仅导致训练时神经网络学习不充分，更导致模型评估可能出现误差。因此十折交叉验证是必须采用，以保证评估模型效果准确的方法。理论上讲，如果能够提供足量的样本进行训练，深度学习方法可以更好的识别微生物与疾病之间的内在联系，从而超越机器学习算法。

其三在于，对于图神经网络而言，输入数据不是图会对模型的分类效果带来很大的影响。由于现有的相关度计算方法以及阈值界定算法仍然难以避免划分过于绝对的问题，从而导致由OTU表生成图时损失了一部分样本特征信息，图神经网络学习图时又损失了一部分样本信息，这可能会导致模型学习能力有限，而无法给出令人满意的结果。

## 2.4 本章小结

本章详细阐释了基于肠道微生物和机器学习预测样本患病与否的方法，共计在六种不同的数据集上评估各种模型，并得到了较好的结果。本章详细论述了数据的来源与构建方式，尤其是图神经网络的数据预处理方式，随后仔细介绍了训练时如何设置参数以追求算法效果最佳化，并且提出了若干方法以改善模型预测能力。结果显示随机森林算法既快又好，可以很高效的在四个数据集上得到最好的效果。这也为后续算法的实际应用提供了殷实的理论支持。

1. **基于随机森林算法的关键物种正常丰度估计算法**

## 3.1 前言

在上一章节中，本研究设计了五种分类器可以用来分析六种疾病与肠道微生物的关联关系并可以据此诊断样本健康情况，并完成了对各种模型性能的评估，为后续应用场景提供了理论依据，如：如何选择最适合的模型来辅助给出诊断结果，如何根据实际应用场合个性化定制模型参数。出于实用性考虑，如果可以给出患者于怎样的丰度下可以恢复健康，将会使本研究的结果更具有参考价值。因此，本章节的目的是从微生物丰度层面给出患者肠道各菌种的调整意见，以试图将患者体内的关键致病菌种调整至正常丰度以使其恢复健康，这样一来将对医学治疗有较大的指导作用。

本章算法受到了上一章中对各种肠道菌群进行了随机森林算法隐含的特征选择，并对其重要度排序得出了一个微生物贡献序列。在优化时发现，排名靠前的微生物被认为与疾病的发作有着难以割舍的关系。迄今为止，大量研究结果都指示着肠道微生物丰度可以反应身体的健康情况，亦或者说影响人体健康情况。本研究基于随机森林算法特征排序的方法构建了评估菌种正常丰度的算法。在之前的分类任务中，机器学习通过学习输入数据寻找数据与标签之间的关系，旨在获得对新的数据进行分类的能力，是一种因果模型。而本算法的初衷是找到合适的输入数据以使得模型可以将数据归为某一类标签，具体而言，微生物的丰度为多少时，样本健康；这样的在已知样本标签类别，寻找契合标签的数据的算法属于逆因果模型。

## 3.2 算法设计

算法具体流程图见图3-1。首先依然采用特征排序算法来选择重要微生物，训练在整个数据集上的随机森林算法按照重要度顺序化微生物丰度，之后从前到后，优先调整排名靠前的微生物争取能够在尽可能少的修改菌群丰度的情况下完成逆因果任务，因为排名靠前的微生物更具备影响最终分类结果的能力。在数据预处理中，所有微生物的丰度都介于一个特定的区间内，在该区间中以特定的步长不断调整微生物的丰度，当步长为0.05时，丰度会在0至1见不断调整如：[0.05,0.1,0.15,…,1]，每次调整丰度后都使用在整个数据集上训练好的分类器上来预测当前健康概率。遍历区间时以能够使得分类器打分最高（健康概率）的丰度为微生物新丰度。这个过程结束后一个微生物的丰度就被确定了，接下来对下一重要菌为算法对象，实施相同的丰度预测流程。值得一提的是如果在本算法中所有尝试的新丰度都不能够提升模型预测为阴性的概率，那么就保持原始丰度。

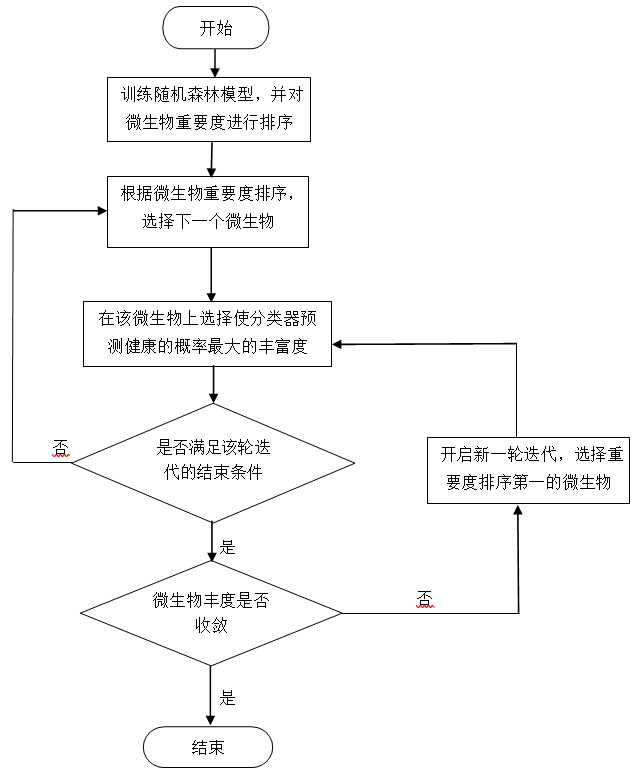


图3-1 逆因果模型算法流程图

由此可见，该算法是一个不断迭代不断更新的过程。迭代设置两个结束条件，满足结束条件时退出迭代。

1. 如果当前轮迭代中调整的微生物数量已经达到最大限制（50），超过这个数值所需要的成本过大，得不偿失，所以当微生物数量极限时停止迭代。

2. 分类器预测样本为健康的概率达到一定数值（75%），超过这个概率以后样本就有较大可能被预测为阴性，本算法认为成功拯救了该病人，遂停止迭代。

可以得出这两个条件是存在一定的矛盾的，因为取得较好的结果需要调整大量微生物，而调整微生物越多预测样本为阴性的概率一定越大。如果不加以限制，模型会得出一个完美样本使得预测为阴性的概率为1，也就是所有的样本都会受到同样的诊疗意见。但是这样的意见纯粹是出于模型自身的学习角度，无法根据每个样本的实际情况，客观公正的定制合理公证的个性化诊疗建议。

对每个微生物的丰度调整迭代将持续进行，直到满足以上条件。此时样本体内的微生物丰度已经足够用来被分类器预测为健康，因此再将其肠道菌群丰度从数据预处理后的特定区间内调整回原值，作为模型给出的诊疗建议，至此算法结束。本系统使用了多线程技术来加速疾病诊断，每一个患病样本都单独开辟了一个线程，这对人性化的给出快速结果十分有意义。

## 3.2 结果与分析

在第二章中分析过，随机森林算法不仅能够快速的给出预测结果，而且也能在多个数据集上取得最好的效果，因此本算法中的分类器选用随机森林分类器。本算法选择的参数为最大微生物数量50，最大健康概率0.7，步长0.05。最终算法的结果如图所示。

为了证明本系统给出的诊疗建议具有实用价值，本研究分析了建议结果。在cirrhosis数据集上，医学界郭栗良子等人的的研究显示该病患者普遍具有显著高的韦荣球菌科和链球菌科细菌，而本研究中给出的建议在大部分患者样本中都指出应该将这两科的微生物丰度降为0，这与医学研究中的结论一致。再比如IBD数据集中，主要的致病菌种毛螺菌科和双歧杆菌也被预测为丰度应降至0，这也符合医学常识。

概言之，本章中构建了基于随机森林算法的逆因果模型算法，算法首先确定关键物种，再调整关键物种的丰度来提高样本被预测为正常的概率，预测结果也在其他研究中有所证实，在一定程度上可以辅助医生给出诊疗建议。建议结果如图3-2所示：图中每两行为一组诊疗建议，一组中的第一行为菌种学名，第二行为具体的丰度修改方案。该算法也会输出样本ID，方便实际使用时对具体的某一样本给出个性化建议。



图3-2 针对某样本的微生物正常丰度估计结果

值得一提的是，虽然诊疗建议部分选用了在评估阶段表现最佳的随机森林算法，但是实际使用时这部分的可参考性与疾病诊断部分选用的分类器也有一定的关系，例如：当使用未经大量训练的GNN作为分类器时，由于模型尚未学习到疾病信息，分类器给出的诊断结果大部分都是阳性，这会直接导致第三部分中的诊疗建议没有用武之地。因此在实际将本算法投入使用时，还应当选择acc较高的模型，以确保给出的建议相对正确。

## 3.3 本章小结

本章基于随机森林算法的特征选择算法以及基于随机森林算法训练的分类器，设计了逆因果算法，旨在找到能够较快的找到使患病样本达到健康水平的微生物丰度，同时一部分预测结果也在其他的研究中得以印证，因此能够对医生的治疗建议提供帮助。

1. **基于机器学习与肠道微生物组的疾病辅助诊断系统的设计与实现**

## 4.1 需求分析

据现有相关学术分析，某些特定的疾病与寄宿在人体内的微生物群落丰度息息相关，而人体内超过80%的微生物都寄宿在肠道之内，作为人体内最大的微生物寄宿场所，肠道微生物丰度以及多样性对人体健康和疾病有着直接的影响[1]

大量由肠道微生物群落导致的疾病的预防以及治愈，在社会经济快速发展的当下，越来越成为人们迫切需要解决的健康性需求，相关领域医学技术的发展也赫然成为了追逐健康时代下的“刚需”，然而肠道微生物群落极其庞大，不进行预测直接对疾病和群落之间的关系进行分析将会消耗难以承受的人力和精力，所以辅助诊断的技术在二者关系的分析领域是不可或缺的。本系统实现的功能恰好可以基于机器学习的方法快速且较为准确的给出诊断结果和诊疗意见，也可以供给医务人员参考，结合丰度建议进行更进一步的医疗指导。

### 4.1.1 系统的业务流程

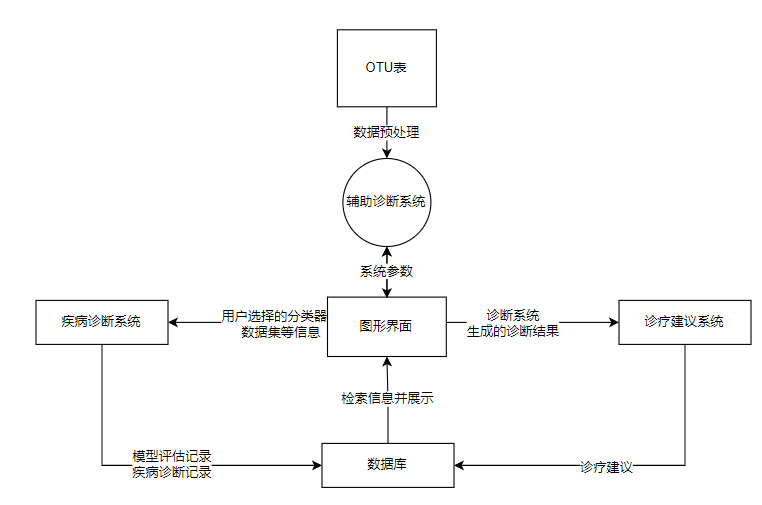


图4-1系统的数据流图

系统的数据流图如图4-1所示，具体数据流包括：

* + 1. 系统读取供后续诊疗使用的OTU表，并进行数据预处理
    2. 将用户从前台界面选择的参数传入疾病诊断系统
    3. 将诊断系统产生的结果传入诊疗建议系统
    4. 将诊断记录和诊疗建议存储进数据库以持久化
    5. 将具体的模型评估诊断记录诊疗建议等展示在前台

### 4.1.2 系统的具体需求

从用户角度看，本系统要实现疾病辅助诊断系统。因此，首先需要实现友好的用户界面方便用户使用，包括模型评估，疾病诊断，记录查看等模块；还需要读取在疾病诊断和模型评估等模块根据用户需求，自定义模型输入参数，包括分类器，数据集，模型具体参数设置等等。

综上所述，系统的具体需求如下：

1. 读取前台界面传入的系统参数，如：分类器，数据集
2. 数据预处理，将被选择到的数据集压缩到特定区间
3. 模型评估，在数据集上以十折交叉验证的方式评估模型
4. 疾病诊断，加载保存好的模型进入系统来进行疾病诊断
5. 根据具体诊断结果给出诊疗建议
6. 数据持久化，将以上功能的执行结果存储在数据库
7. 给出参数设置建议，针对某种疾病将在数据库中表现最好的模型参数返回前台用户界面。

### 4.1.3 用例分析

本章系统的用户是医护人员，医护人员需要进行模型评估，疾病诊断，查看历史记录等操作，结合以上需求，系统具体的用例分析图如图4-2所示：

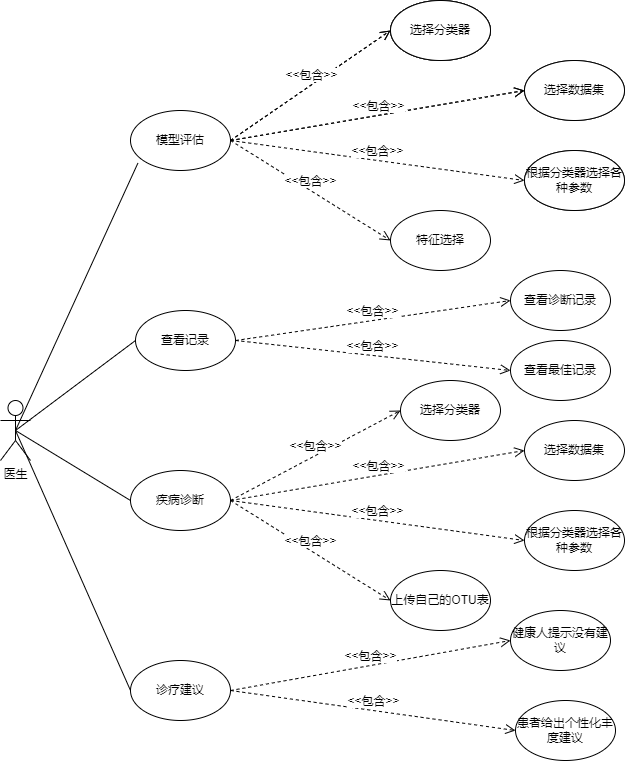


图4-2用例分析图

（1）模型评估时，可以选择分类器数据集等参数配置，具体的用例描述如表4-1所示。

表4-1医生模型评估用例描述

|  |  |
| --- | --- |
| 描述项 | 说明 |
| 用例名称 | 模型评估 |
| 参与者 | 医生 |
| 前置条件 | 医生登录进入辅助诊断系统系统，点击进入模型评估界面 |
|  |  |
| 基本事件流 | 1. 医生进入系统初始界面 2. 医生点击登录，进入系统操作界面 3. 医生选择分类器数据集等信息 4. 医生点击模型评估后给出结果 |

（2）查看记录时，可以选择分类器数据集等参数配置，具体的用例描述如表4-2所示。

表4-2医生模型评估用例描述

|  |  |
| --- | --- |
| 描述项 | 说明 |
| 用例名称 | 查看记录 |
| 参与者 | 医生 |
| 前置条件 | 医生完成了模型评估或者疾病诊断操作，生成了具体记录和结果 |
|  |  |
| 基本事件流 | 1. 医生完成了模型评估或者疾病诊断 2. 医生进入查看记录界面，可以根据分类器数据集，acc等指标排序 3. 模型评估记录中可以查看最优记录，即模型表现最佳时的效果 4. 疾病诊断记录中，可以查看诊断时的参数设置，点击查看详情可以查看具体某次诊断中全部样本的结果 |

（3）疾病诊断时，可以选择分类器数据集等参数配置，具体的用例描述如表4-3所示。

表4-3医生模型评估用例描述

|  |  |
| --- | --- |
| 描述项 | 说明 |
| 用例名称 | 疾病诊断 |
| 参与者 | 医生 |
| 前置条件 | 医生登录进入辅助诊断系统系统，点击进入疾病诊断界面 |
|  |  |
| 基本事件流 | 1. 医生进入系统初始界面 2. 医生点击登录，进入系统操作界面 3. 医生选择分类器数据集等信息 4. 医生上传待诊断样本的OTU表 5. 医生点击开始诊断，诊断完成后提示前往诊疗结果界面 |

（4）查看诊疗建议时，医生进入系统具体诊疗记录界面，点击查看详情可以查看到具体的某一病人的诊断结果，再点击查看详情可以查看到该具体病人的诊疗建议，如表4-4所示。

表4-4医生查看诊疗建议用例描述

|  |  |
| --- | --- |
| 描述项 | 说明 |
| 用例名称 | 查看诊疗建议 |
| 参与者 | 医生 |
| 前置条件 | 医生完成了疾病诊断操作，点击进入具体诊疗结果界面，点击查看诊疗建议 |
|  |  |
| 基本事件流 | 1. 医生进入具体诊疗结果界面 2. 医生点击查看诊疗建议 3. 医生进入具体诊疗建议界面 4. 医生点击某一菌种名称查看具体丰度修改方案 |

## 4.2 概要设计

### 4.2.1 功能模块划分

系统分为评估记录查看，模型评估，疾病诊断，诊断结果查看模块。功能模块图如图4-3所示。在模型评估模块，用户主要根据自己的个性化需求来定制分类器和数据集，以评估各个模型的性能来作为后续使用的指导。在疾病诊断模块，用户可以采用最佳的模型来进行诊断，也可以个性化的定制诊断模型来诊断疾病。最终在诊断结果查看模块，用户不仅可以查看到所有样本的患病情况，分类器情况等信息，还可以查看到系统根据用户个体情况定制的诊疗建议，具体为将某种菌从某丰度调整到某丰度。

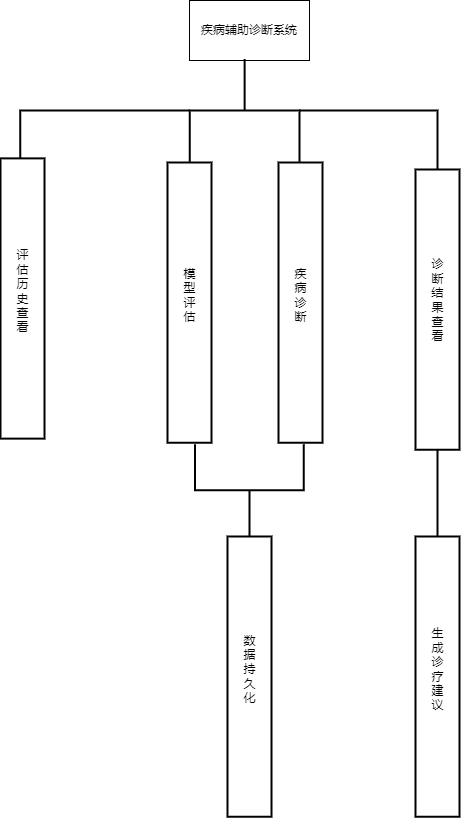


图4-3 系统功能模块图

### 4.2.2 数据库设计

由于每次评估模型时都会生成一条记录，为了高效的将记录从后台读取到前台，本系统中使用数据库来存储模型评估记录以及疾病诊断记录，方便操作员查看。本着轻便快捷的原则，选择sqlite作为本系统的数据库，数据库中共设计了四个表，分别是：

1.EvaluatlonHistory表用来存储评估记录，该表的具体构成如下表4-5所示

表4-5 EvaluatlonHistory表结构

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 限制 | 描述 |
| HistoryID | int | PRLMARY KEY AUTO\_LNCREMENT | 主键 |
| dataset | text | NOT NULL | 数据集 |
| classifier | text | NOT NULL | 分类器 |
| feature | text | DEFAULT '' | 选择前多少重要的菌 |
| epoch | int | NOT NULL | 训练轮数 |
| batchsize | int | NOT NULL | 批大小 |
| threshold | int | DEFAULT '' | GNN生成用的阈值 |
| acc | text | NOT NULL | 训练的accuracy |
| auc | text | NOT NULL | ROC曲线下面积 |
| date | text | CURRENT\_TLMESTAMP | 评估完成时间 |

2.DlagnosisHistory表用来存储诊断记录，主要记录诊断基本信息，该表的构建方式和约束如下表4-6所示，该表的HistoryID会被用作后续表的外键之一。

表4-6 DlagnosisHistory表结构

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 限制 | 描述 |
| HistoryID | int | PRLMARY KEY AUTO\_LNCREMENT | 主键 |
| dataset | text | NOT NULL | 数据集 |
| classifier | text | NOT NULL | 分类器 |
| feature | text | DEFAULT '' | 选择前多少重要的菌 |
| epoch | int | NOT NULL | 训练轮数 |
| batchsize | int | NOT NULL | 批大小 |
| threshold | int | DEFAULT '' | GNN生成用的阈值 |
| date | text | CURRENT\_TLMESTAMP | 评估完成时间 |

3.SpecflcDlagnosis表用来存储某一次诊断记录中所有样本的具体诊断信息，该表的构建方式和约束如下表4-7所示，该表的SampleID会被用作后续表的外键之一。

该表有表级约束：PRLMARY KEY(“SampleID”,“HistoryID”)，由于某一个数据集可能被多次诊断，所以不能单独使用SampleID作为主键，必须要细化为某一次诊断中的某一个具体病人才能确保唯一性。

表4-7 SpecificDiagnosis表结构

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 限制 | 描述 |
| SampleID | text | NOT NULL | 主键之一 |
| Sex | text | NOT NULL | 数据集 |
| Age | text | NOT NULL | 分类器 |
| Country | text | DEFAULT '' | 选择前多少重要的菌 |
| Result | text | NOT NULL | 训练轮数 |
| HistoryID | int | FORELGN KEY REFERENCE DiagnosisHistory(‘HistoryID’) | 外键关联到上一个表，同时作为主键之一 |

4.DieaseTreatment表用来存储某一次诊断记录中某一样本的具体诊疗建议，该表的构建方式和约束如下表4-8所示。该表有表级约束：PRLMARY KEY(“Name”,“SampleID”,“HistoryID”)，类似的由于多次诊断中可能会出现同名菌种，必须要细化为某一次诊断中的某一个具体病人肠道中某一种菌的丰度建议才能确保唯一性。

表4-8 DieaseTreatment表结构

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 限制 | 描述 |
| Name | text | NOT NULL | 细菌的学名 |
| Treatment | text | NOT NULL | 丰度建议，例如（0.001->0.123） |
| SampleID | text | REFERENCE SpecificDiagnosis (‘SampleID’) | 外键关联到上一个表，同时作为主键之一 |
| HistoryID | int | FORELGN KEY REFERENCE DiagnosisHistory(‘HistoryID’) | 外键关联到其他表，同时作为主键之一 |

## 4.3 详细设计

系统主要由三个模块组成，首先是模型评估与疾病诊断模块，在这个模块中医疗人员可以自由的组合五种分类器和六种疾病，并根据每种分类器的特征设置参数，例如：DNN和1D-CNN可以设置训练轮次epoch，批大小batchsize，而GNN可以设置相关度划分时的阈值。同时模型评估时还可以使用特征筛选的方法，只针对重要微生物进行计算。。

在模型评估界面中，系统操作员选择参数并运行后会显示：正在评估模型，请耐心等待。待后台评估结束后，前台会显示评价结果，包括acc和auc，同时系统会生成一条模型评估记录，详细记录参数设置，数据集，分类器，评估效果和评估时间等，并将该数据插入数据库的评估记录表中，以备后续查看结果。在评估记录界面中，本系统还具有存储和展示最佳模型的功能，以方便操作员参考，在何种设置下的分类器最具备提供合理诊疗建议和诊断结果的能力。

接下来是疾病诊断模块，在该模块中，系统拥有类似模型评估界面的布局，操作员同样可以自由的组合五种分类器和六种疾病，并根据每种分类器的特征设置参数。包括进行特征选择筛选重要微生物等等。值得注意的是该界面中设置了预置模型，操作人员可以简单的在预置模型菜单中选择数据集，该操作将会读取前端被选择中的数据集，并进入数据库中在模型评估记录表上查找在该数据集上评估结果最好的模型，读取其参数至前台界面以供操作员选择。其中如果选用图神经网络进行疾病诊断时，系统将会在诊断疾病时不断将该数据集上表现最好的图神经网络与当前训练中的模型进行比较，如果本次训练的结果要好于之前训练结果则将最佳结果替换为当前训练中的模型。否则在进行诊断时仍然使用在该数据集上效果最好的模型进行诊断，以尽可能地避免错诊现象。

最后一个模块是诊断记录模块，该模块集成了查看诊断记录，查看诊断结果，查看诊疗建议三者于一身。用户进入该界面后可以看到在第二个模块运行时生成的诊断记录，记录了使用的分类器，数据集，参数设置以及诊断时间。查看详情后将进入数据库中的具体记录表，以诊断记录id为key搜索该表中与该次诊断相关的全部记录，这些记录分别对应参与诊断的每一个样本，随后将这些结果返回到前台界面。在前端用户可以看到具体的样本id，国籍，年龄，性别等信息。同时每个样本都有一个查看诊疗建议按钮，其中被诊断为健康的样本被点击时会提示：“健康人没有诊疗建议”。而点击患病样本时，会在数据库的诊疗建议表中以样本id，诊断记录id为keys搜索具体的针对某一个患者的诊疗建议，该记录中包含多种菌及与其一一对应的丰度修改方案。返回前台界面后用户可以查看具体的诊疗措施，从而结合医学技术分析具体病人并且给出方案。

## 4.4 系统的实现与测试

本系统的算法部分和后端部分完全由python实现，算法部分的具体实现在第二章中已经做过介绍，后端部分使用flask框架部署在本地服务器的5000端口和5555端口，前端使用vue框架实现，界面使用element ui美化，前后端数据交互使用axios实现。IDE使用Visual Studio Code安装各种插件运行，在python3.9，windows 10环境下运行。

在各个模块的测试中，本测试不关注模型的具体效果，因为已经在第二章中进行过了详尽的阐释，本章节重点关注测试系统能否正确运行并得到结果。本测试共采用了以下用例，结果涵盖了系统的三大模块，以及每次操作后记录表的变化。具体验证见图4-4至图4.4(p)。

1. 模型评估测试：

用户可以自由的选择各种分类器，并且跳帧分类器参数，点击开始评估后，提示正在评估，如图所示。



图4-4 评估时选择svm分类器

评估结束后，系统将评估结果返回前台，显示本次评估的acc和auc，并将该记录插入数据库中。



图4-5 svm分类器的评估结果

系统的本模块选择不同分类器可以设置不同的参数，很大程度上满足了用户的个性化定制需求。

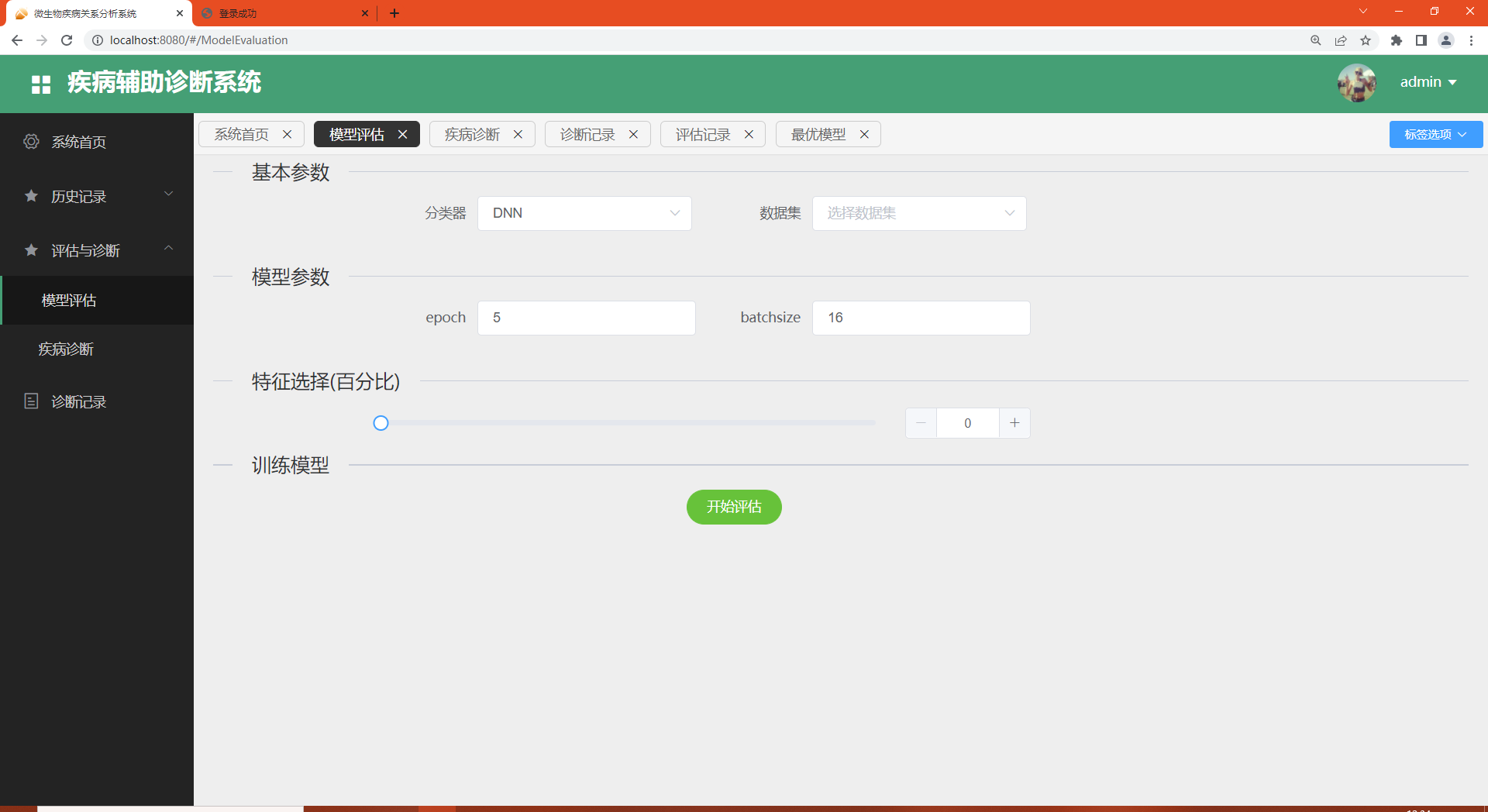


图4-6 评估时选择dnn分类器

当某必填参数没有填写时，系统会提示，参数值不能为空。

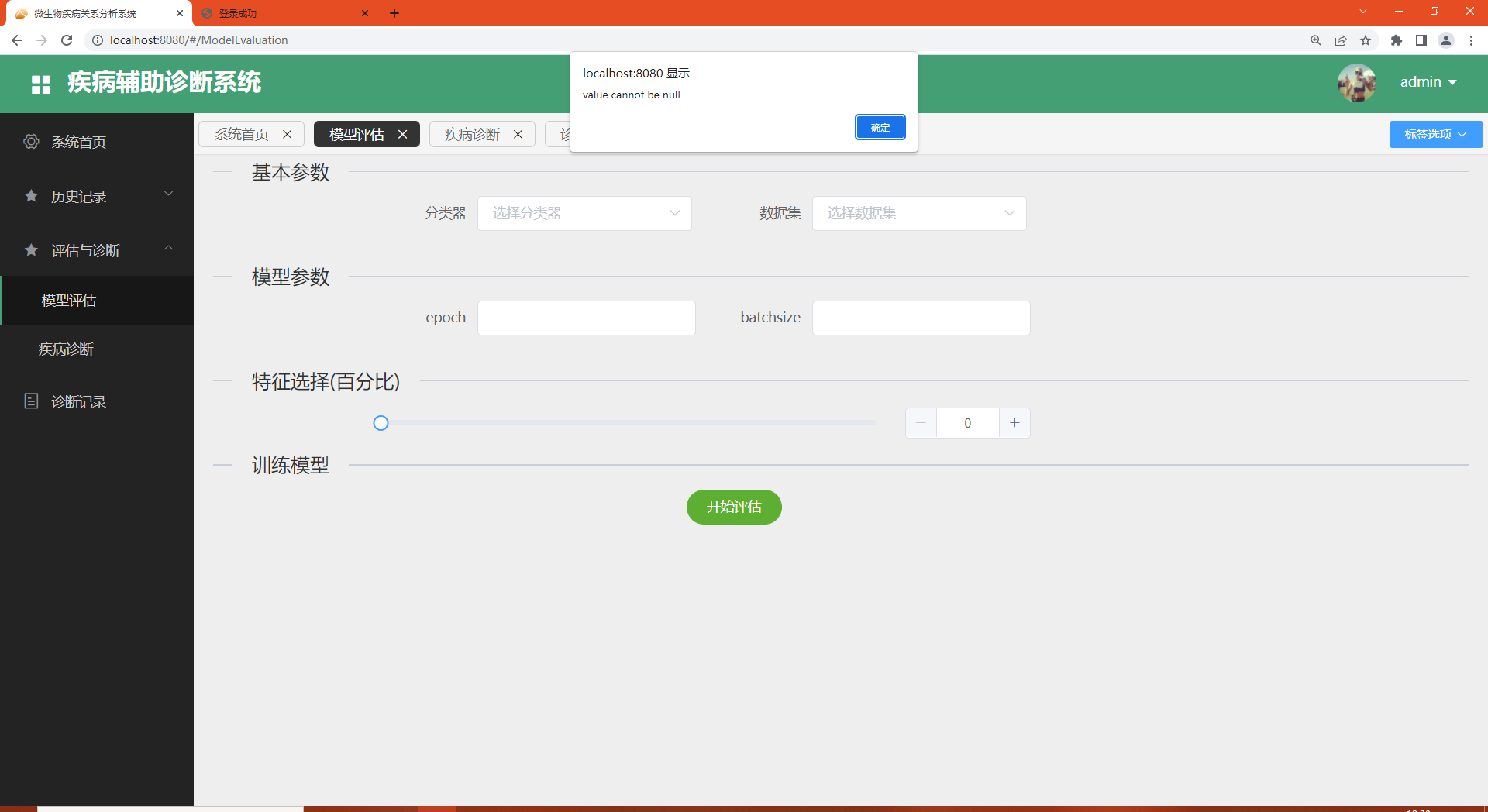


图4-7 评估参数为空时的提示

当完成一次模型评估后，系统会将诊疗记录插入数据库，以方便后续查看，该记录可以实时同步在系统中。

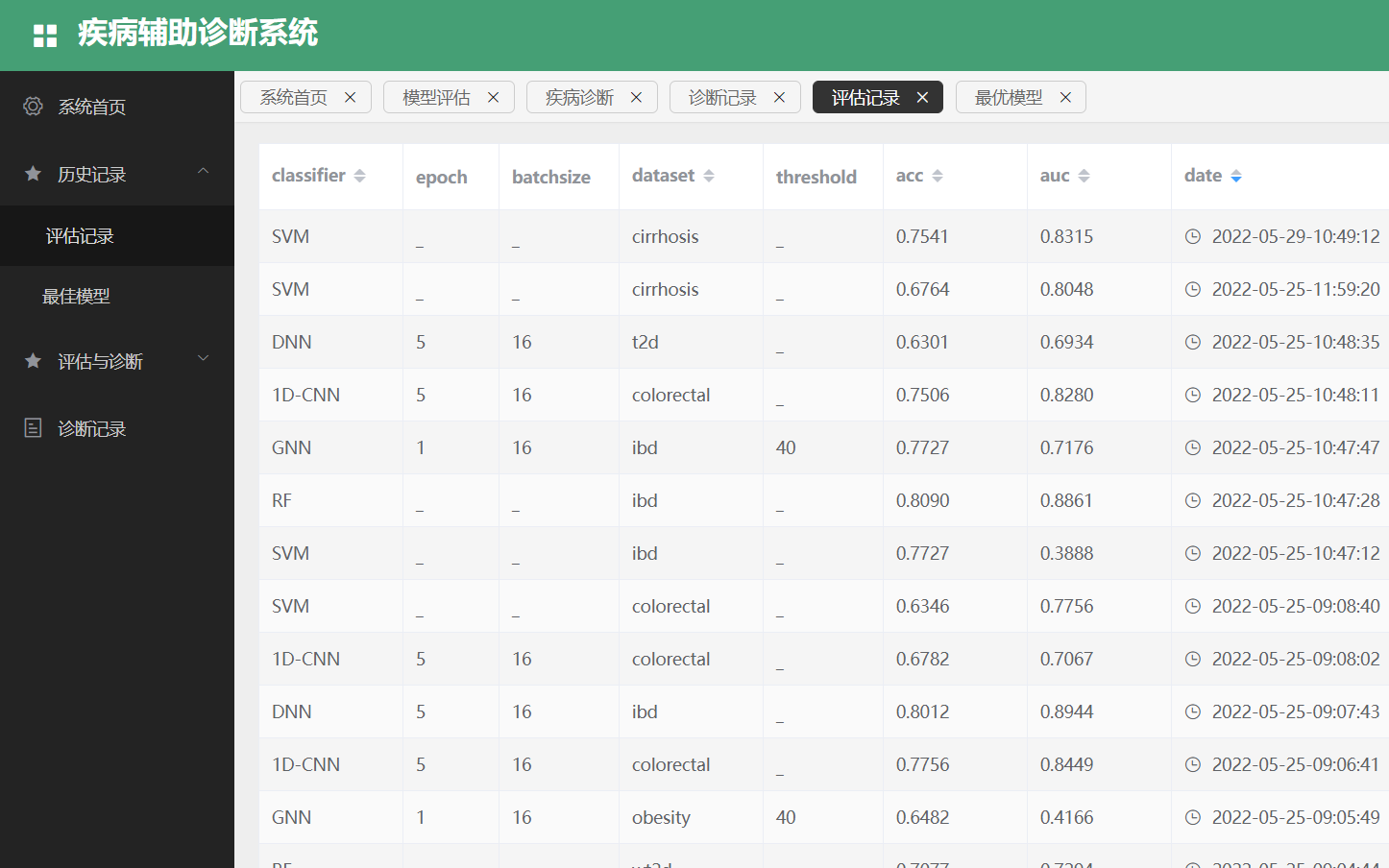


图4-8 评估完成后生成记录

如果模型的表现超过了最佳模型中的表现，那么最佳模型中的记录将会被替换为当前模型。

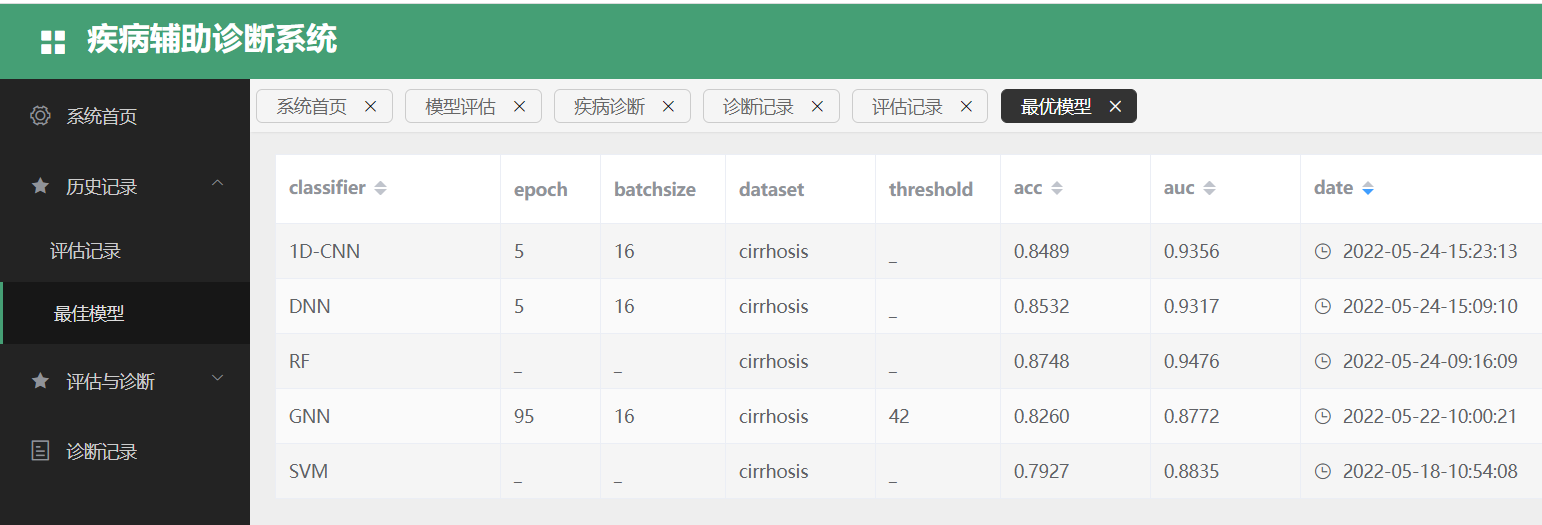


图4-9 最佳记录

为了测试系统的鲁棒性，本研究还额外测试了图神经网络GNN的评估能力和记录，由于训练轮数少，用例耗时为30.74秒，主要时间用来生成微生物相互作用网络。同时该用例中采用了特征选择的方法，以测试系统优化方法的实现效果。



图4-10 评估模型时进行特征选择

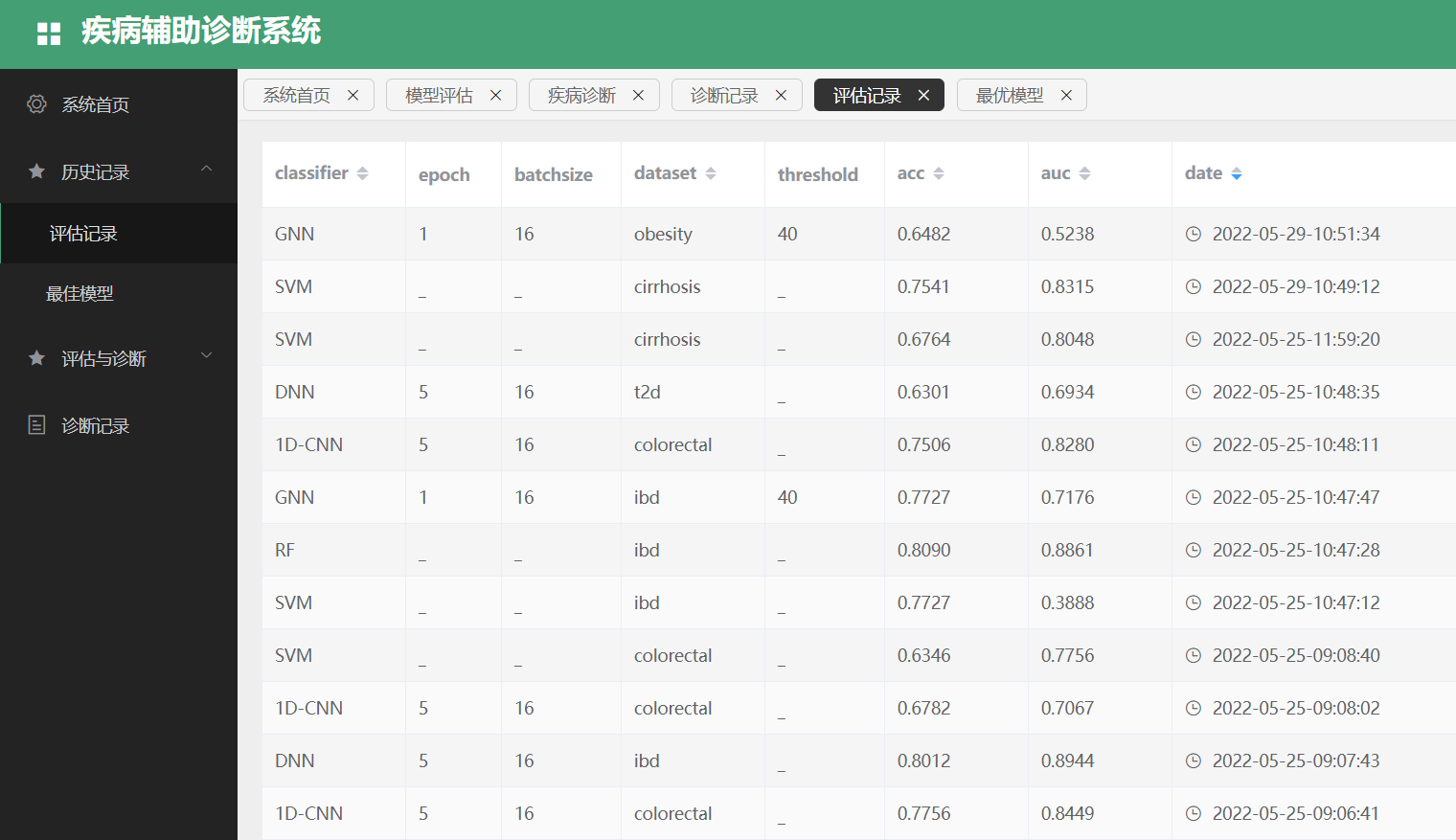


图4-11 特征选择下的结果展示

在疾病诊断界面，启用预置模型时会自动的读取数据库中，表现最佳模型的各项参数设置，并将这些参数填写在前端界面中，方便医生参考使用。

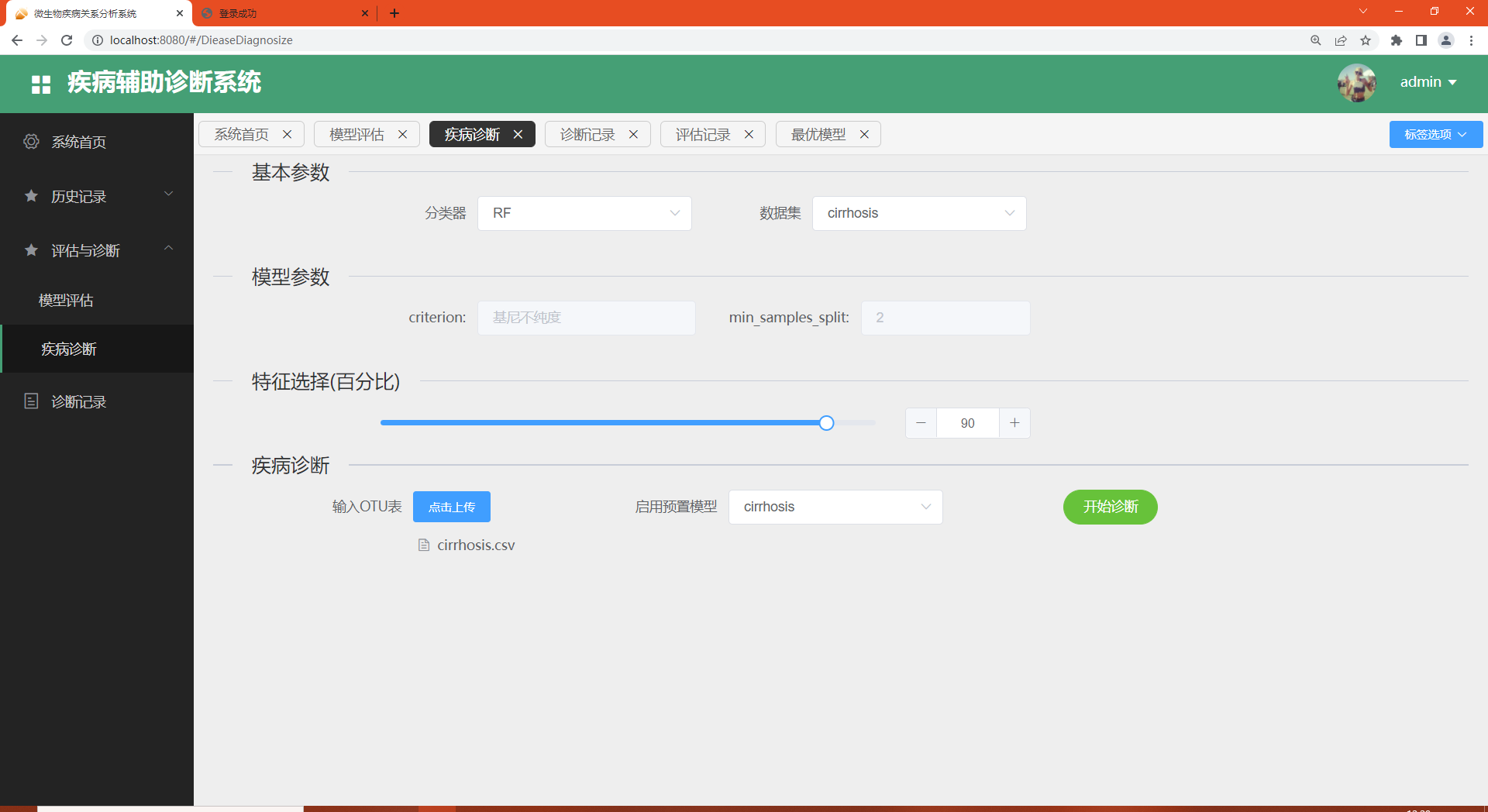


图4-12 启用预置模型

在疾病诊断界面试图上传多个OTU表时会提示出错，因为诊断时只用到一个OTU表。

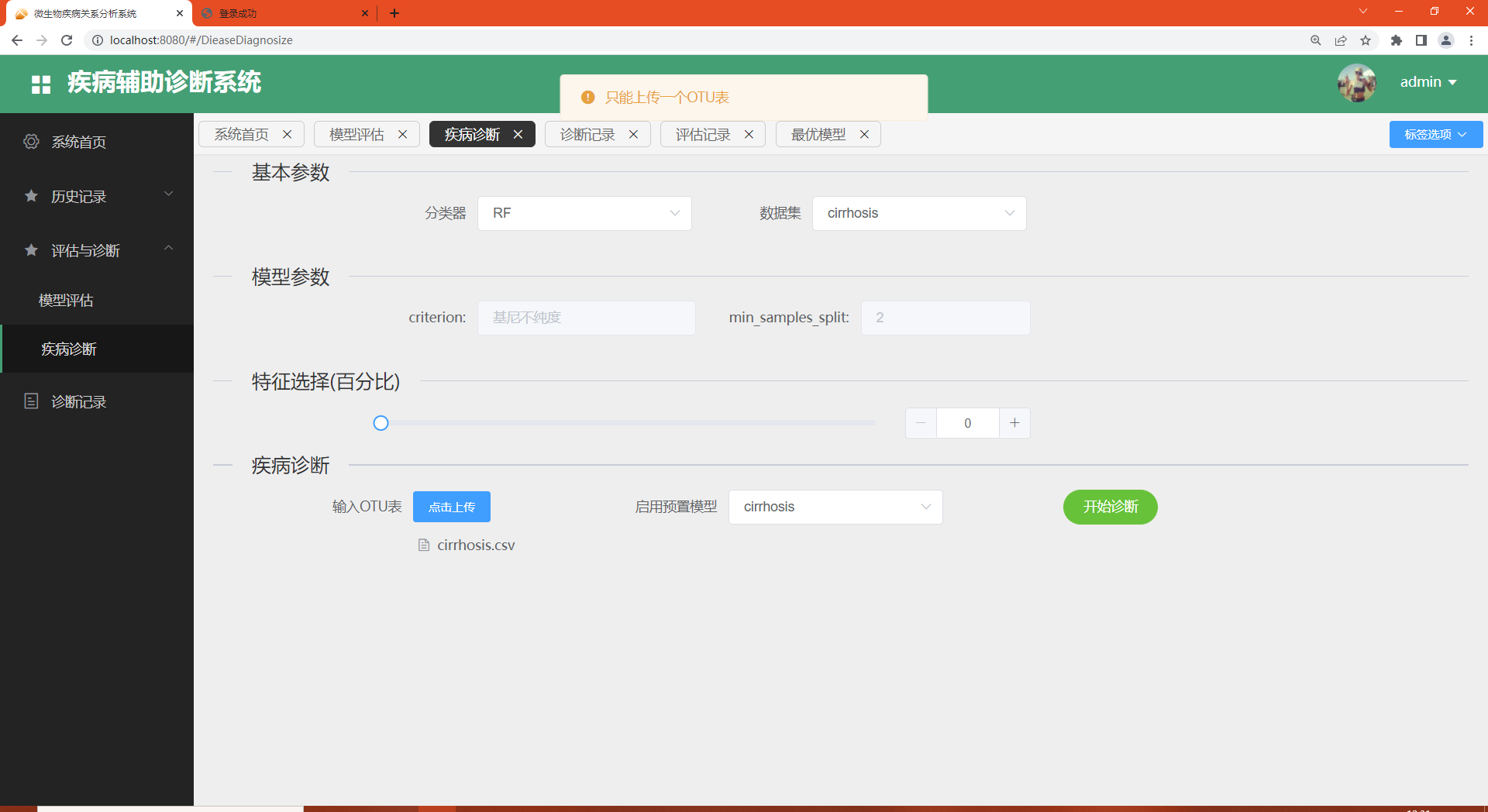


图4-13 疾病诊断时试图上传多个OTU表时提示

在疾病诊断界面，选择好参数，上传好OTU表后，点击开始诊断，会提示正在诊断，诊断时间漫长，请耐心等待。

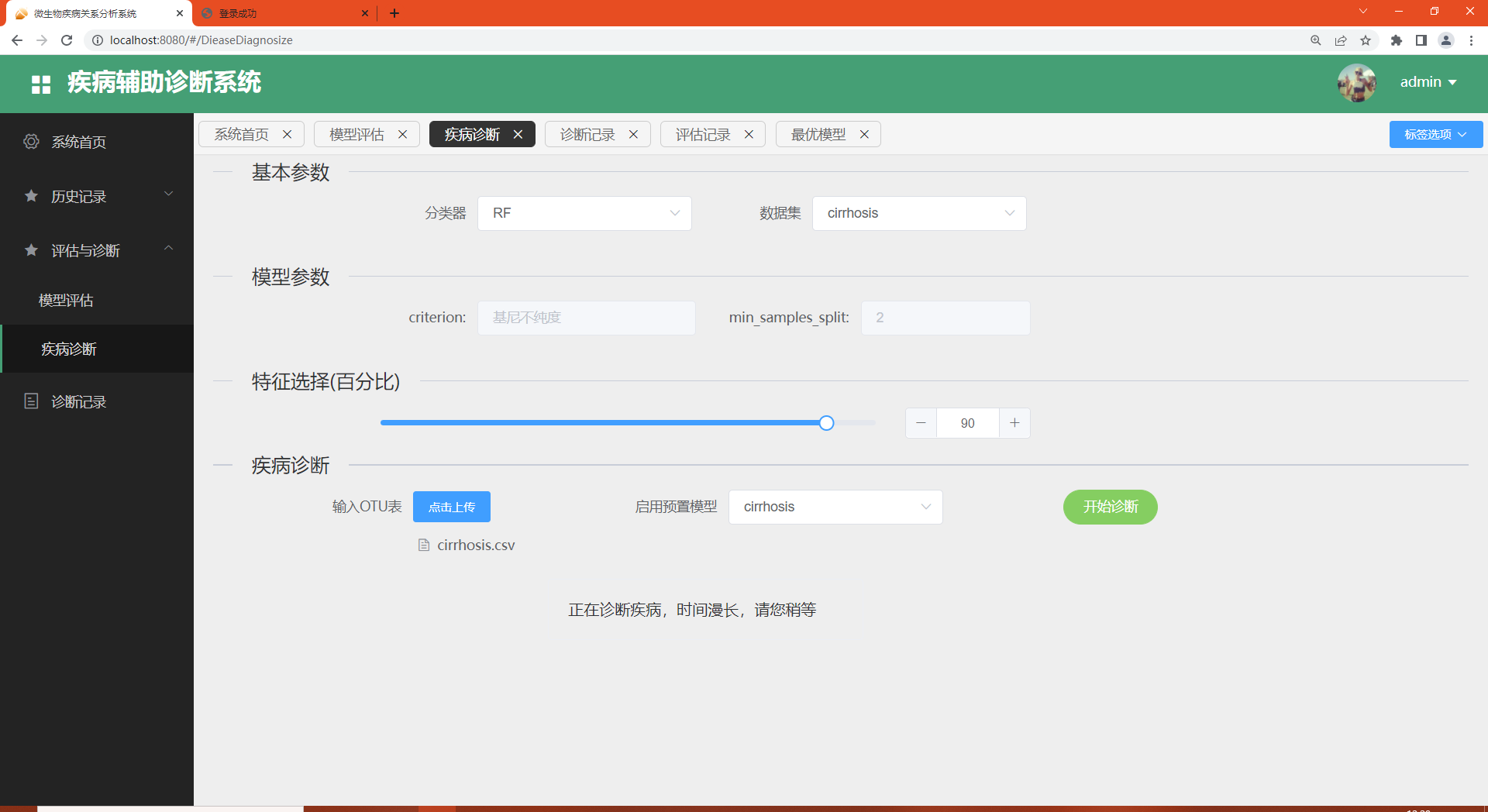


图4-14 诊断疾病时提示信息

当疾病诊断结束后，系统提示诊断结果已经记录在了诊断记录界面，可以前往查看

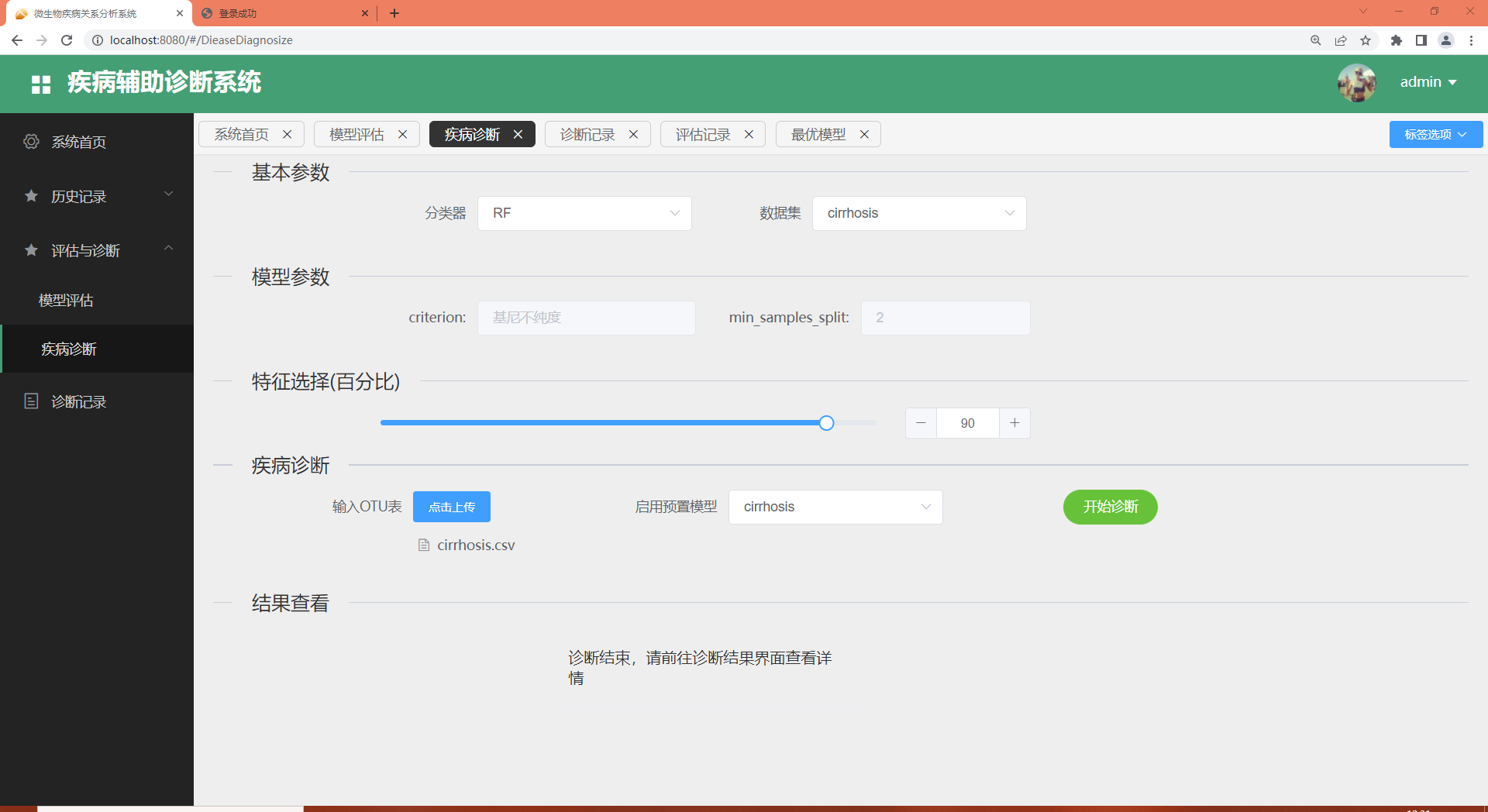


图4-15诊断结束后提示前往诊断记录界面查看结果

在诊断记录界面，可以查看到每次进行疾病诊断时的参数设置，诊断时间，分类器数据集等信息。

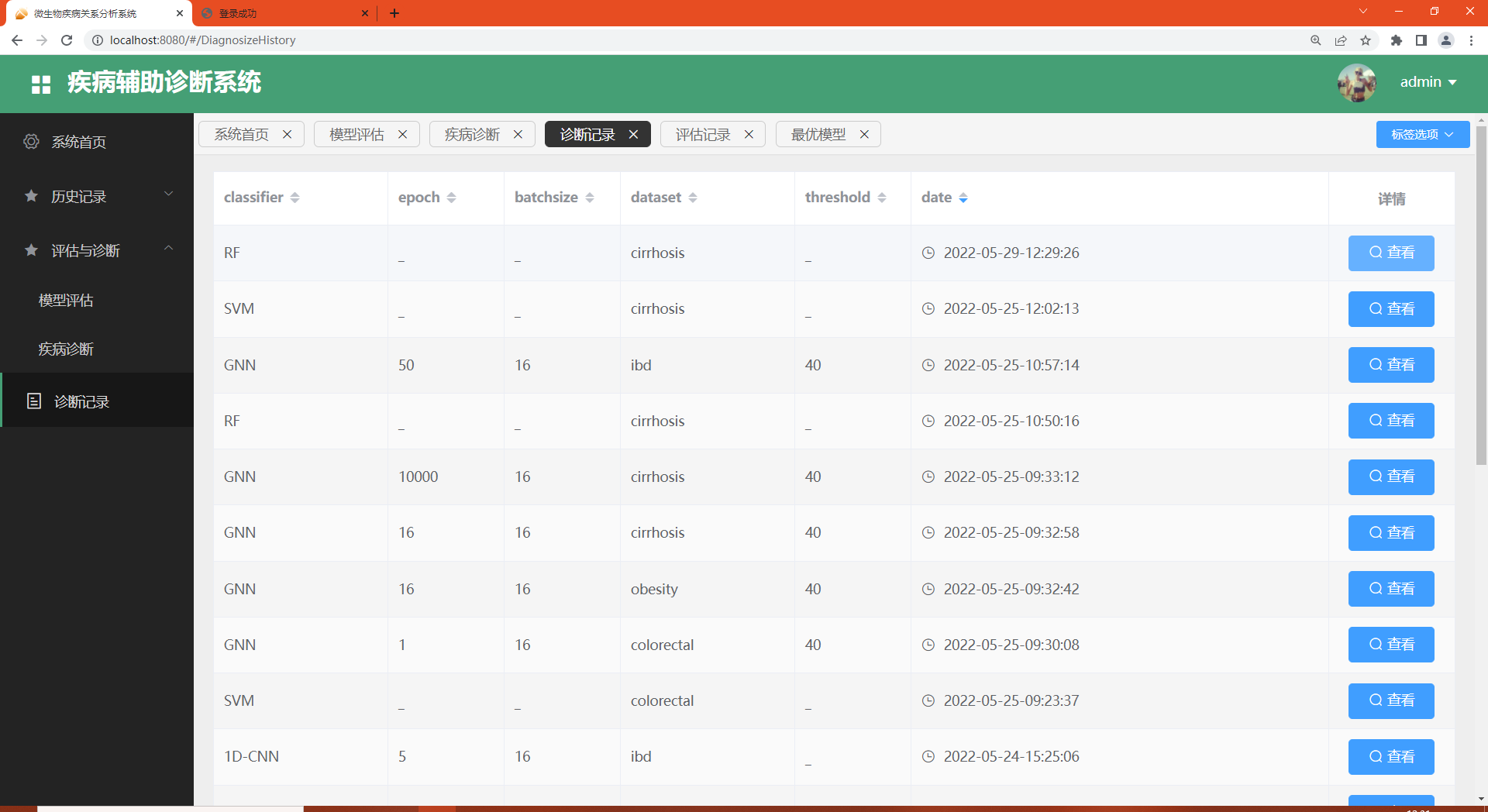


图4-16诊断记录查看

点击诊断记录中某一条记录最后的按钮后，会进入具体结果界面，可以查看这次诊断结果中所有病人的年龄，性别，国籍等信息以及诊断结果

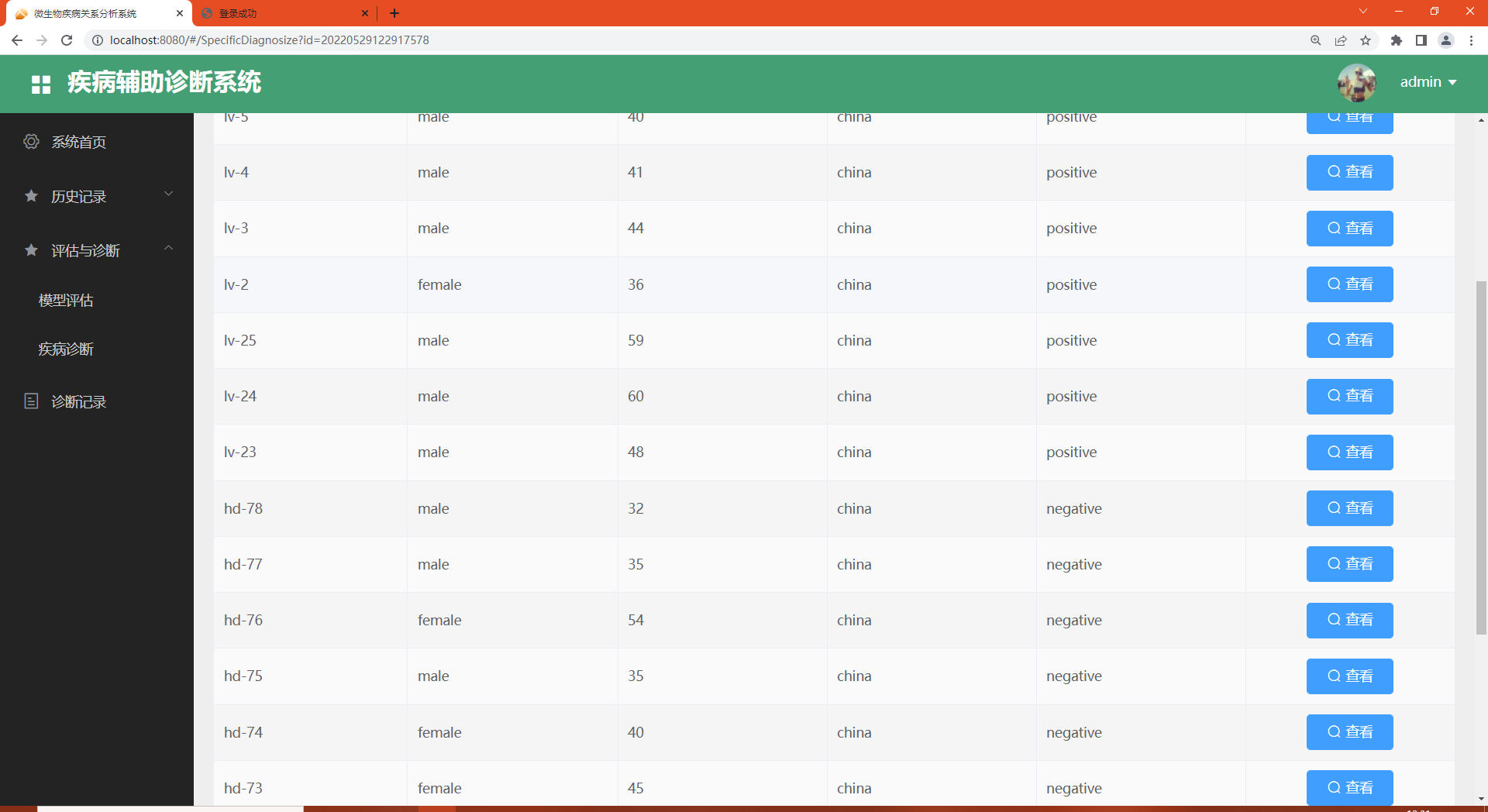


图4-17某次诊断记录的具体结果展示

在此界面点击查看按钮，可以查看为每一位病人给出的个性化诊疗建议，但是点击健康人时，会提示健康人没有诊疗建议。

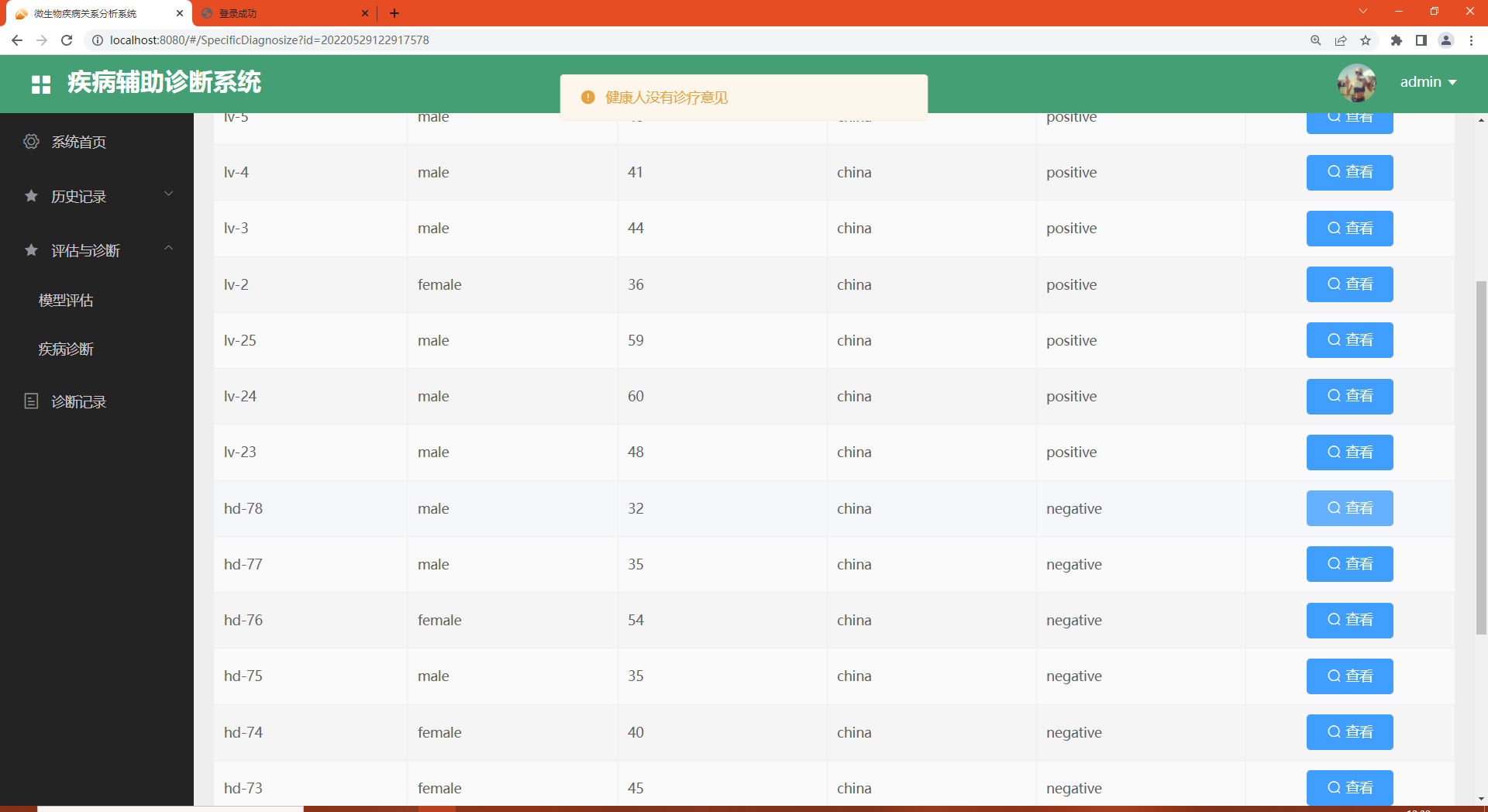


图4-18系统提示健康人不需要诊疗建议

进入患病病人界面可以查看被系统建议修改丰度的关键物种名，展开某一物种后可以查看到给出的具体丰度建议。

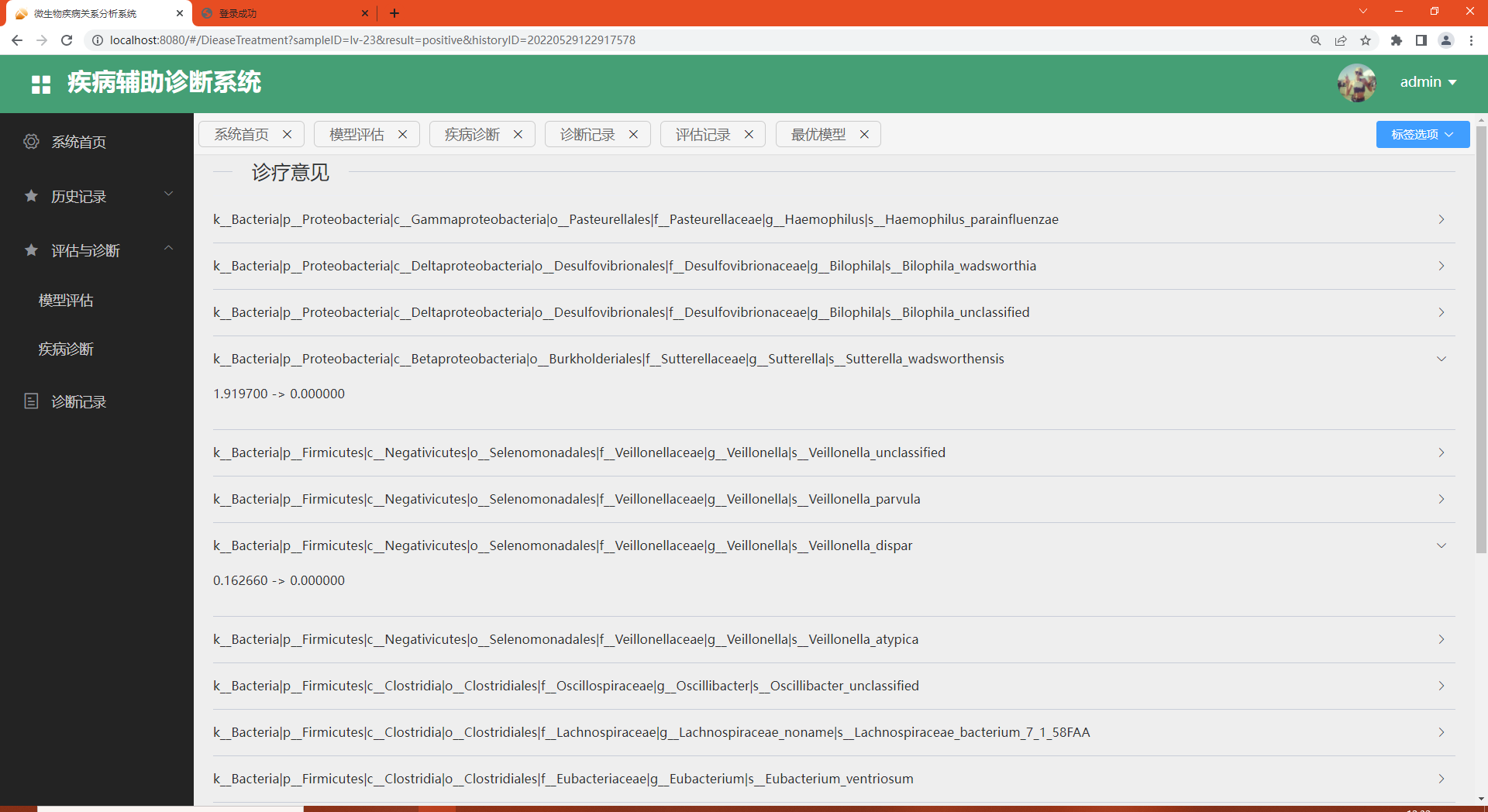


图4-19患者的具体丰度建议

以上即为本系统的全部系统设计，在经过完整的测试后，功能全部可以正常运行，可以作为疾病辅助诊断系统投入使用。

## 4.5 本章小结

本章首先分析了用户需求，从系统业务流程出发，分析了系统的具体需求，并且站在医生的角度进行了用例分析，以确保本系统能满足医生的使用需求。随后本章出发于设计的角度，首先进行了系统的概要设计，包括功能模块划分和数据库设计，以方便后续详细设计和实现。接下来本章详细的设计了各个具体的功能模块，以确保系统能够涵盖从评估模型到诊断建议的全部需求。最终，本章详细的测试了系统所有的功能模块，以确保可以作为疾病辅助诊断系统投入使用。

**结 论**

本研究的主要工作是将五种分类器应用在六种常见与肠道微生物相关的疾病数据集上进行疾病诊断，目前关于肠道微生物与疾病的主流研究方向是预测某些疾病与某些微生物之间是否存在关系，也就是预测在所有人体内疾病和微生物之间的关系而不是针对某些具体样本。而本研究从肠道微生物微丰度出发，使用机器学习和深度学习方法直接预测样本患病与否，同时也借助微生物丰度构建了微生物相互作用网络，并且将新颖的图神经网络应用于本研究中，并且在IBD数据集上的表现超越了随机森林算法。其次，为了提高神经网络预测的效果，除了根据分类器调整参数和应用适合本研究的模型以外，本研究主要尝试了两种方法，即特征选择方法，以及数据集交叉扩充方法，从结果角度，这些方法能够改善模型的分类效果。从最终结果来看，在 IBD 数据集和Obesiy数据集上深度学习结果较好，而在其余 4 个数据集上机器学习方法效果好而且推理快。因此，本研究认为随机森林算法作为机器学习方法的代表，可以高效准确的诊断样本健康与否。

本研究的创新点有以下两方面：其一是本研究将新颖的图神经网络中的图分类任务应用于疾病诊断领域，据本人所知，此前没有人尝试过这项工作，虽然取得的结果并不尽好，但是可以为未来的相关工作提供指引和导向。其二是本研究可以根据样本自身特征给出个性化的诊疗建议，这对于疾病辅助诊断而言有着一定的参考价值，使得本系统更加实用。

未来的工作主要是寻找更加合适的构建肠道微生物的方法，以尽可能地避免错误地构建微生物相互作用网络，从而影响后续图分类任务的准确率。此外，图神经网络中Graph Attention Networks虽然不能很好的应用于本研究，但是他能够提取相互作用网络中的注意力。将提取出的注意力作为辅助，加入到全连接神经网络中，也就是每个节点与下一层链接时引入基于注意力的权重，或许可以在一定程度上改善深度学习方法的效果。最后，本研究最大的桎梏仍是数据，目前公开的可用于疾病分类的肠道微生物数据集较少，随着测序技术的发展日后可能会涌现更多能体现样本特征的数据集，届时模型在经过大量训练后效果或许可以改善。

**本科毕业论文（设计）原创性声明**

本人呈交给哈尔滨工业大学的学位论文，除所列参考文献和世所公认的文献

外，全部是本人毕业设计期间在导师指导下的研究成果。除文中已经标明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

若有不实之处，本人愿意承担相关法律责任。

作者签名：

导师签名：

年 月 日

**参考文献**

[1] Methé BA, Nelson KE, Pop M, Creasy HH, Glgllo MG, Huttenhower C, Gevers D, Petroslno JF, Abubucker S, Badger JH. A framework for human mlcroblome research. Nature. 2012;486(7402):215.

[2] Consortlum HMP. Structure, functlon and dlverslty of the healthy human mlcroblome. Nature. 2012;486(7402):207–14.

[3] Nothnagel EA, McNell M, Albershelm P, Dell A. Host-pathogen interactlons. Plant Physlol. 1983;71(4):916-26.

[4] Medzhltov R. Recognltlon of mlcroorganlsms and actlvatlon of the lmmune response. Nature. 2007;449(7164):819–26.

[5] Tlaskalova-Hogenova H, Stepankova R, Kozakova H, et al. The role of gut mlcroblota (commensal bacterla) and the mucosal barrler ln the pathogenesls of lnflammatory and autolmmune dlseases and cancer: contrlbutlon of germ-free and gnotoblotlc anlmal models of human dlseases. Cell Mol Lmmunol, 2011, 8(2): 110-20

[6] Mallnen E, Rinttllä T, Kajander K, et al. Analysls of the fecal mlcroblota of lrrltable bowel syndrome patlents and healthy controls wlth real-tlme PCR [J]. Am J Gastroenterol, 2005, 100: 373−382.

[7] Qln J, Ll Y, Cal Z et al . A metagenome wide assoclatlon study of gut mlcroblota ln type 2 dlabetes. Nature 2012;49 0( 74 18)18):5 5 60.

[8] Zhang W, Jlang S, Qlan DW, et al. Metabollsm of narlngln produced by intestlnal bacterla [J]. Acta Pharm Sln (药学学报), 2013, 48: 1817−1822.

[9] anssen AW, Kersten S. The role of the gut mlcroblota ln metabollc health [J]. FASEB J, 2015, 29: 3111−3123.

[10] Sharma Dlvya; Xu Wel phyLoSTM: a novel deep learnlng model on dlsease predlctlon from longltudlnal mlcroblome data. [J]Applled Blochemlstry and Mlcroblology,2021

[11] doa rdo Pasolli, Duy Tln Truong, Falzan Mallk et al. Machine Learnlng eta analysls of Large Metagenomlc Data sets: Too ls and Blologlcal Lnslghts 2016, 12(7)

[12] M anandha r Lshan, Allmadadl Ahm ad, A ryal Sachln et al Machine Learnlng of Gut Mlcroblome Compo sltlon for Dl agnostlc Classlflcatlon of Lnflammatory Bow el Dlseases[J]. The FASEB Journal, 2021,35

[13] Yahul Long, Jlawel Luo, Yu Zhang, Yan Xla, Predlctlng human mlcrobe–dlsease assoclatlons vla graph attentlon networks wlth lnductlve matrlx completlon, Brleflngs ln Blolnformatlcs, Volume 22, Lssue 3, May 2021, bbaa146.

[14] Huang, YA., You, ZH., Chen, X. et al. Predlctlon of mlcrobe–dlsease assoclatlon from the integratlon of nelghbor and graph wlth collaboratlve recommendatlon model. J Transl Med 15, 209 (2017).

[15] Qu J, Zhao Y and Yln J (2019) Identlflcatlon and Analysls of Human Mlcrobe-Dlsease Assoclatlons by Matrlx Decomposltlon and Label Propagatlon. Front. Mlcroblol. 10:291. dol: 10.3389/fmlcb.2019.00291

[16] J. Luo and Y. Long, "NTSHMDA: Predlctlon of Human Mlcrobe-Dlsease Assoclatlon Based on Random Walk by Integratlng Network Topologlcal Slmllarlty," ln LEEE/ACM Transactlons on Computatlonal Blology and Blolnformatlcs, vol. 17, no. 4, pp. 1341-1351, 1 July-Aug. 2020, dol: 10.1109/TCBB.2018.2883041.

[17] 吴桐,王鸿超,陆文伟,等.肥胖人群肠道菌群特征分析及机器学习模型[J].微生物学通报,2020,47(12):4328-4337.

[18] Deng, Y., Jlang, YH., Yang, Y. et al. Molecular ecologlcal network analyses. BMC Blolnformatlcs 13, 113 (2012).

[19] Ruan, Q., D. Dutta, M. S. Schwalbach, J. A. Steele, J. A. Fuhrman and F. Sun (2006). "Local slmllarlty analysls reveals unlque assoclatlons among marlne bacterloplankton specles and envlronmental factors." Blolnformatlcs 22(20): 2532-2538.

[20] Vellkovl P , Cucurull G , Casanova A , et al. Graph Attentlon Networks[J]. 2017.

[24] Deng Y, Jlang Y H, Yang Y, et al. Molecular ecologlcal network analyses[J]. BMC blolnformatlcs, 2012, 13(1): 1-20.

[25] Reshef D N, Reshef Y A, Flnucane H K, et al. Detectlng novel assoclatlons ln large data sets[J]. sclence, 2011, 334(6062): 1518-1524.

[26] 游懿君,梁丹丹,陈天璐.4种相关分析方法在菌群和代谢物相关研究中的初步比较[J].转化医学杂志,2018,7(02):93-96.

[27] Defferrard M, Bresson X, Vandergheynst P. Convolutlonal neural networks on graphs wlth fast locallzed spectral fllterlng[J]. Advances ln neural lnformatlon processlng systems, 2016, 29.

[28] Hamllton W, Ylng Z, Leskovec J. Lnductlve representatlon learnlng on large graphs[J]. Advances ln neural lnformatlon processlng systems, 2017, 30.

[29] Vellckovlc P, Cucurull G, Casanova A, et al. Graph attentlon networks[J]. stat, 2017, 1050: 20.

**致 谢**

时光飞逝，大学四年的生活转瞬即逝。从大一的青涩懵懂，到大四的懵懂。细细想来，身上少了几分稚气后，就不再有那么多稚气了。不过四年的时光总该有些收获，感谢大学四年里遇到的老师们，以认真严谨的态度对待每一位学生的作业和考试，让我能够将“规格严格，功夫到家”的校训作为对自己的训诫。感谢身边沉迷内卷的一位位卷王同学们，让我意识到学海无涯苦作舟的真谛。感谢徐永东老师，在我渴望技术的时候给了我加入实验室的机会，并信任的派给我重要的工作，闲暇之余，徐老师甚至愿意亲自前往实验室与我约谈，帮我为未来的学术生涯出谋划策。徐老师在毕设期间时常与我联系沟通难处解决问题，也能够包容我因为升学而不得不一心三用。