分类号：　　密　　级：公开

学　号：20222508002　　单位代码：10759

石河子大学
硕 士 学 位 论 文

基于集成学习与群智能优化算法的糖尿病诊断方法研究与系统实现

学 位 申 请 人　刘如如

指 导 教 师　徐丽萍 教授

刘伟 副教授

申请学位门类级别　专业硕士

学科、专业名称　电子信息

研 究 方 向　计算机技术

所 在 学 院　信息科学与技术学院

中国·新疆·石河子

2025 年 6 月

分类号：　　密　　级：公开
学　号：20222508002　　单位代码：10759

石河子大学
硕 士 学 位 论 文

基于集成学习与群智能优化算法的糖尿病诊断方法研究与系统实现

学 位 申 请 人　刘如如
指 导 教 师　徐丽萍 教授
　　刘伟 副教授

申请学位门类级别　专业硕士
学科、专业名称　电子信息
研 究 方 向　计算机技术
所 在 学 院　信息科学与技术学院

中国·新疆·石河子
2025 年 6 月

Research and System Implementation of Diabetes Diagnosis Method Based on Ensemble Learning and Swarm Intelligence Optimization Algorithms

A Dissertation Submitted to

Shihezi University

In Partial Fulfillment of the Requirements

for the Degree of

Master of Engineering

By

Liu Ruru

Electronic Information

Dissertation Supervisor: Prof. Xu Liping and Prof. Liu Wei

June, 2025

石河子大学学位论文独创性声明及使用授权声明

学位论文独创性声明

本人所呈交的学位论文是在我导师的指导下进行的研究工作及取得的研究成果。据我所知，除文中已经注明引用的内容外，本论文不包含其他个人已经发表或撰写过的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中作了明确的说明并表示谢意。

研究生签名：  时间： 2025 年 5 月 26 日

使用授权声明

本人完全了解石河子大学有关保留、使用学位论文的规定，学校有权保留学位论文并向国家主管部门或指定机构送交论文的电子版和纸质版。有权将学位论文在学校图书馆保存并允许被查阅。有权自行或许可他人将学位论文编入有关数据库提供检索服务。有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

研究生签名：  时间： 2025 年 5 月 26 日

导师签名：  时间： 2025 年 5 月 26 日

摘要

糖尿病的诊断在现代医疗中至关重要，作为健康管理的关键部分，充分挖掘医疗大数据以优化糖尿病诊断算法具有重要意义。本文深入研究并优化糖尿病诊断算法，旨在提高诊断的准确性，为临床决策提供可靠支持。本文的主要研究内容包括以下几个方面：

（1）基于改进沙猫群优化算法的糖尿病特征选择方法研究。首先，介绍了糖尿病诊断所用的数据集，包括数据来源、特征描述以及数据预处理方法。接着，采用改进的沙猫群优化算法（MSCSO）对糖尿病数据进行特征选择，目的是筛选出与糖尿病诊断最相关的特征。在 D1 数据集中，MSCSO 将特征数量减少了约 67.66%；在 D2 数据集中，特征数量减少了约 67.35%。通过与原始数据集、互信息法、方差选择法以及卡方过滤法进行对比，实验结果表明，MSCSO 在特征选择方面具有显著优势，能够有效提升糖尿病诊断的准确性与效率。

（2）基于优化的 Stacking 集成学习的糖尿病诊断算法研究。首先提出了一种基于分类错误率权重策略的 Stacking 集成学习改进方法，并设计了改进的麻雀优化算法（ISSA）。通过斯皮尔曼相关系数分析不同机器学习模型之间的关联程度，选取了 LR、DT、KNN 和 XGBoost 作为基学习器。经过实验分析，最终选择 SVM 作为元学习器，并成功构建了糖尿病诊断模型。为了进一步提高算法性能，使用 ISSA 对模型进行超参数优化。与原始的 Stacking 集成学习模型相比，优化后的 Stacking 集成学习模型在 D1 数据集上，Accuracy、Precision、Recall 和 F1 Score 分别提升了 2.75%、2.99%、3.05%和 2.92%；在 D2 数据集上，分别提升了 1.97%、1.54%、2.43%和 1.93%。相较于改进后的 Stacking 集成学习模型，优化后的 Stacking 集成学习模型在 D1 数据集上的四个指标分别提升了 1.51%、1.18%、1.38%和 1.73%；在 D2 数据集上，分别提升了 1.32%、1.50%、1.10%和 1.29%，实验结果表明，优化后的 Stacking 集成学习模型有效提高了糖尿病诊断的准确性和可靠性。

（3）糖尿病诊断系统的设计与实现。系统采用 Django 框架作为后端处理数据存储与模型调用，前端使用 Vue.js 构建，确保良好的用户体验。系统集成了优化后的 Stacking 集成学习模型，进一步提升了诊断准确率。主要的功能模块包括用户管理、诊断管理和体检数据管理，旨在提供一个高效、直观的糖尿病诊断平台。

关键词：糖尿病诊断；机器学习；集成学习；群智能优化算法

Abstract

The diagnosis of diabetes is crucial in modern medicine. As a key part of health management, it is of great significance to fully mine medical big data to optimize diabetes diagnosis algorithms. This thesis delves into the research and optimization of diabetes diagnosis algorithms, aiming to improve diagnostic accuracy and provide reliable support for clinical decision-making. The main content of this study includes the following aspects:

(1) Research on feature selection method for diabetes based on an improved sand cat swarm optimization Algorithm. First, the data set used for diabetes diagnosis is introduced, including the data source, feature description, and data preprocessing methods. Then, the improved sand cat swarm optimization (MSCSO) algorithm is applied to diabetes data for feature selection, aiming to select the most relevant features for diabetes diagnosis. In the D1 dataset, MSCSO reduced the feature count by about 67.66%, and in the D2 dataset, the feature count decreased by about 67.35%. Compared to the original dataset, mutual information method, variance selection method, and chi-square filtering method, experimental results show that MSCSO has significant advantages in feature selection and can effectively improve the accuracy and efficiency of diabetes diagnosis.

(2) Research on diabetes diagnosis algorithm based on optimized Stacking ensemble learning. This section proposes an improved Stacking ensemble learning method based on a classification error rate weighting strategy and designs an improved sparrow search algorithm (ISSA). Spearman's rank correlation coefficient is used to analyze the correlation between different machine learning models, and LR, DT, KNN, and XGBoost are selected as base learners. After experimental analysis, SVM is chosen as the meta-learner, and a diabetes diagnosis model is successfully constructed. To further enhance the algorithm's performance, ISSA is used to optimize the model's hyperparameters. Compared to the original Stacking ensemble model, the optimized Stacking ensemble model shows improvements in Accuracy, Precision, Recall, and F1 Score by 2.75%, 2.99%, 3.05%, and 2.92%, respectively, on the D1 dataset; and by 1.97%, 1.54%, 2.43%, and 1.93%, respectively, on the D2 dataset. Compared to the improved Stacking ensemble model, the optimized Stacking ensemble model improves the four metrics by 1.51%, 1.18%, 1.38%, and 1.73% on the D1 dataset, and by 1.32%, 1.50%, 1.10%, and 1.29% on the D2 dataset. The experimental results show that the optimized Stacking ensemble model significantly enhances the accuracy and reliability of diabetes diagnosis.

(3) Design and implementation of diabetes diagnosis system. The system uses the Django framework for backend data storage and model invocation, with the frontend built using Vue.js to ensure a good user

experience. The system integrates the optimized Stacking ensemble learning model, further improving diagnostic accuracy. The main functional modules include user management, diagnostic management, and physical examination data management, aiming to provide an efficient and intuitive diabetes diagnosis platform.

Key words: Diabetes Diagnosis; Machine Learning; Ensemble Learning; Swarm Intelligence Optimization Algorithm.

目录

第 1 章 绪论

1.1 研究背景与意义

在信息化社会的推动下，全球医疗数据量呈现出爆炸式的增长[1-3]。现代医疗技术的进步使得健康数据的收集和存储变得更加高效和便捷，从而积累了大量的健康信息。这些数据不仅包括患者的生理指标、病史记录，还涵盖了各种实验室检测结果。随着数据量的激增，如何从这些海量的信息中提取有用的知识，辅助医生进行疾病诊断，成为当前医疗行业面临的重大挑战[4-7]。在此背景下，数据驱动的智能诊断技术被广泛引入医疗领域，通过挖掘数据中的潜在模式，显著提升疾病诊断的精准度与时效性。特别是在糖尿病等慢性疾病的诊断中，如何利用这些数据成为提高诊断水平的关键。

基于医疗大数据的广泛应用，国家层面也陆续出台了与慢性病相关的政策文件来引导医疗行业的科技创新[8-9]。《“健康中国 2030”规划纲要》中明确指出，到 2030 年中国要将主要慢性病的过早死亡率降低 30%，糖尿病被列为优先防治的慢性病之一[10]。此外，根据国家卫生健康委员会发布的《慢性病防治工作规划（2021-2025）》，通过加强糖尿病的筛查和早期诊断，力争将糖尿病患者的控制率提升至 50%。这些政策不仅为糖尿病的管理提供了强有力的支持，也为研究和应用新兴的诊断系统提供了政策保障。

尽管政策支持为糖尿病的管理提供了框架，但糖尿病的发病率依然持续上升[11-13]。根据 2023 年国际糖尿病联盟（IDF）的报告，全球已有约 5.37 亿成年人受糖尿病影响，预计到 2045 年这一数字将增加到超过 7 亿[14-16]。中国作为全球糖尿病负担最重的国家之一，成人糖尿病患病率已达到 11.2%。糖尿病的高发不仅对个人健康造成严重影响，也对医疗系统和社会经济带来巨大的压力。糖尿病患者常伴随心血管疾病、肾病和视网膜病变等并发症，这些长期并发症进一步加重了医疗负担[17]。因此，提高糖尿病的诊断效率和准确性，对于减轻患者的医疗负担和降低社会经济成本具有重要意义。这种背景下，糖尿病诊断系统的研究显得尤为迫切。

糖尿病诊断系统作为交叉学科的创新实践，可通过多维度健康数据的深度解析，显著提升糖尿病诊断的效率和准确性。通过对大量健康数据的综合分析，该系统能够帮助医生更准确地识别糖尿病患者，减少误诊和漏诊的风险。这不仅改善了诊断过程，还能够优化医疗资源配置，特别是在基层医疗机构中，弥补了医生资源不足的缺陷。此外，它可以提升糖尿病的早期诊断率，减少患者的医疗负担，并为未来医疗智能化的发展奠定基础。整体来看，这一研究方向不仅符合国家推动医疗智能化的战略目标，还能为实现健康中国战略提供重要支持，推动糖尿病管理水平的整体提升。

1.2 国内外研究现状

随着人工智能技术的发展，糖尿病诊断广泛应用了机器学习相关方法。本文将相关研究分为三类：传统机器学习方法、集成学习方法和群智能优化方法。其中，集成学习是机器学习的重要分支，强调模型融合以提升性能；群智能优化方法则常用于特征选择和参数调优，辅助提高模型的准确性与稳定性。

1.2.1 基于传统机器学习的糖尿病诊断研究现状

一些研究人员采用机器学习方法对糖尿病进行诊断，旨在通过自动化的数据分析和模式识别提升诊断效率和准确性。例如，崔波等人[18]提出一种改进的 K 近邻算法（K-Nearest Neighbors，KNN）对 2 型糖尿病患者进行高效诊断，构造 KNN 分类器时使用主成分分析（Principal Component Analysis，PCA）对每个特征赋予不同的权重。刘恬宁等人[19]探究了决策树（Decision Tree，DT）、支持向量机（Support Vector Machine，SVM）和 PCA 等常见的机器学习模型在糖尿病诊断下的进展。杨光等人[20]利用 DT 建立 2 型糖尿病诊断模型，为能更准确地诊断 2 型糖尿病提出理论依据，取得了良好的诊断效果。凌雄娟等人[21]对 DT、逻辑回归（Logistic Regression，LR）、XGBoost、随机森林（Random Forest，RF）、KNN、神经网络 6 种模型的诊断性能进行分析，发现采用 LR 模型在糖尿病诊断的二分类问题上具有较高的稳定性和准确率，XGBoost 模型的召回率最优。通过数据特征分析进一步得出糖尿病的患病率与血糖浓度的相关性最大，与皮肤厚度不相关。郭金旦等人[22]通过比较 LR、SVM、DT、朴素贝叶斯和 KNN 在 2 型糖尿病风险诊断中的性能，发现 LR 在准确性和稳定性方面表现最佳，为临床诊断模型的评估和算法选择提供了重要参考。孟敏敏等人[23]通过队列研究构建了妊娠糖尿病患者产后血糖异常的风险诊断模型，比较了多因素 LR 和 RF 的诊断效果，发现随机森林模型在准确率、精确度、召回率、F1 得分和 AUC 等方面均优于 LR 模型，显示了其在诊断妊娠糖尿病患者产后糖代谢异常风险方面的优越性。怀莹莹等人[24]成功构建了一个基于妊娠早期临床资料的妊娠期糖尿病（Gestational Diabetes Mellitus，GDM）诊断模型，该模型通过 LR 分析识别了多个 GDM 的影响因素，并以此建立了诊断公式。该模型在 ROC 曲线下面积达到 0.818，灵敏度为 93%，特异度为 56%，显示出较好的诊断效能，为早期 GDM 筛查提供了有效的工具。谢妮妮等人[25]利用机器学习算法对糖尿病数据进行分析，通过交叉验证优化模型参数，并以 F1 Score 作为评价标准，以识别糖尿病的主要影响因素并提高诊断准确性。李树华等人[26]通过分析河北省 1,738 名孕妇数据，建立了一个诊断 GDM 的模型，该模型通过单因素和多因素 LR 分析筛选出独立危险因素，

并使用 ROC 曲线评估其诊断效能，结果显示模型拟合优度好，诊断价值高，能有效辅助 GDM 的早期识别和管理。Khanam 等人[27]使用机器学习相关模型对糖尿病进行诊断，并发现 LR 和 SVM 在诊断糖尿病方面表现较好。Krishnamoorthi 等人[28]提出了一个智能机器学习基础的糖尿病诊断架构，该框架使用了多种机器学习技术，如 DT、RF 和 SVM，并在实际数据上进行了评估，结果表明该框架实现了 83%的准确率，并且具有最低的错误率。Ahmed 等人[29]提出了一种融合机器学习方法的糖尿病诊断模型，该模型结合了 SVM 和人工神经网络（Artificial Neural Network，ANN）。模型的诊断结果被用作模糊逻辑模型的输入，从而最终确定糖尿病的诊断结果。这个融合模型的诊断准确率达到了 94.87%，高于以前的方法。Sivaranjani 等人[30]使用了 SVM 和 RF 算法来诊断糖尿病。通过特征选择和 PCA 来提高诊断准确性。结果显示，随机森林的诊断准确率为 83%，高于支持向量机的 81.4%。Rastogi 等人[31]提出了一种糖尿病诊断模型，采用逻辑回归、SVM、随机森林等方法。实验结果显示，逻辑回归模型表现最佳，准确率达到 82.46%。

尽管基于机器学习的糖尿病诊断模型在提高诊断效率和准确性方面展现了良好的前景，但这种方法也存在一些不可忽视的缺点。机器学习模型对特征的选择和数据质量非常敏感，容易受到噪声数据的影响，导致诊断结果不稳定。此外，传统机器学习模型通常依赖于预先定义的规则和参数，可能在处理复杂高维数据时表现欠佳，难以应对多样化的患者数据。

1.2.2 基于集成学习的糖尿病诊断研究现状

为了提高糖尿病诊断的准确性和可靠性，研究者们越来越多地采用集成学习方法，通过结合多个模型的优势，提升了诊断系统的整体性能。例如，刘巧红等人[32]使用 Kaggle 平台的糖尿病数据，利用 XGBoost 构建了糖尿病分类诊断模型，同时引入了 SHAP 增强了模型的可解释性，识别关键的影响特征。周建华等人[33]采用 Stacking 策略对 SVM、CatBoost、XGBoost 进行算法融合。实验证明，融合算法在准确率，精度率，容错率方面都有大幅提高，能够更有效地辅助医生进行糖尿病诊断和干预。马金龙等人[34]的研究，开发了一种妊娠期糖尿病智能诊断系统，通过比较 10 种机器学习模型并使用 Stacking 算法进行集成，系统能够有效诊断妊娠期糖尿病风险，并提供相关健康建议。周乐明等人[35]通过基于 XGBoost、LightGBM、AdaBoost 和多层感知机（Multilayer Perceptron，MLP）等 4 种分类器进行糖尿病的诊断，初步解决了糖尿病早期筛查的问题，可作为一种早期诊断的工具。吴晖南等人[36]成功开发了基于 LightGBM 的高效糖尿病诊断模型，通过优化超参数和数据校正技术，实现了 97%的综合准确率，为糖尿病早期诊断提供了有力的临床决策工具。刘静乐等人[37]通过提出随机森林-交叉验证递归特征消除法（Random Forest - Recursive Feature Elimination with Cross-Validation，RF-RFECV）和

LightGBM 的混合算法，在中国健康与养老追踪调查数据集上实现了 0.9772 的准确率和显著的性能提升，证明了该算法在早期识别糖尿病高危人群和辅助临床诊断中的高效性和准确性。曹长玲等人[38]通过应用统计学和机器学习方法，成功建立了针对糖尿病视网膜病变合并症的诊断模型，并通过 LR 和 LightGBM 算法的比较，证明了 LightGBM 模型在诊断糖尿病视网膜病变合并冠心病、肾病和下肢动脉病变方面具有更高的准确性和诊断效能，为临床预防性治疗提供了重要的参考依据。冯鑫磊等人[39]提出了一种基于妊娠早期体检和基因信息的集成学习方法，设计了改进的 Stacking 集成学习模型，通过基分类器的自适应选择和元层 GBDT 模型的组合学习，显著提升了妊娠期糖尿病的诊断准确性和稳定性，F1 值分别比单一模型和传统 Stacking 模型提高了约 9%和 7%。李佳思等人[40]利用 XGBoost 算法在糖尿病数据集上建立了高准确率的诊断模型，准确率达到 77.83%，并通过 SHAP 模型揭示了葡萄糖浓度、BMI 和年龄是糖尿病诊断的关键因素。Hasan 等人[41]构建了多种机器学习模型，包括 KNN、DT、RF 等，使用加权集成技术，将这些基础的机器学习模型的诊断结果进行融合。通过为每个模型分配不同的权重，提高了整体诊断的准确性和稳定性。Mahesh 等人[42]提出了一种基于集成学习的糖尿病诊断方法，结合了多种机器学习模型。研究显示，该方法显著提高了糖尿病诊断的准确性，为早期诊断提供了有效的新工具。Atif 等人[43]提出了一种基于集成学习的强投票分类器的鲁棒模型，该模型结合了 LR、DT 和 SVM 等多种机器学习算法，该模型在 Pima Indians 糖尿病数据集上达到了 81.17%的准确率，而在早期糖尿病风险诊断数据集上则达到了 94.23%的准确率，显示了其在糖尿病诊断中的优越性能。Ganie 等人[44]提出了一种基于集成学习的框架，用于早期诊断 2 型糖尿病。该方法结合了 Bagging、Boosting 和 Voting 等技术，实验结果表明，Bagging 在准确率、精确度和召回率等指标上表现优异。

综上所述，集成学习在糖尿病诊断中展现了显著优势。通过融合多个模型的预测结果，集成学习不仅提高了诊断的准确性和稳定性，还增强了对复杂数据的处理能力，克服了单一模型的局限性。然而，这种方法也面临挑战，尤其是在超参数选择方面，不恰当的超参数设置可能会影响模型的整体性能。

1.2.3 群智能优化算法在糖尿病诊断中的研究现状

群智能优化算法在糖尿病诊断中越来越受到关注，例如，邹琼等人[45]通过构建麻雀优化算法（Sparrow Search Algorithm，SSA）优化的 BP 神经网络模型，在糖尿病肾病（Diabetic Nephropathy，DN）的早期诊断中表现出色，准确率达到 95.83%，F1-score 高达 0.9600，显著优于其他传统机器学习模型，为 2 型糖尿病患者提供了一种高效、准确的 DN 诊断工具。汪敏等人[46]提出了一种改进的遗传算法（IGABP）优化 BP 神经网络的方法，通过改进选择算子和自适应调整交叉及变异概率，成功构建了糖尿病并发症

诊断模型。实验结果表明，IGABP 在诊断准确率和网络收敛速度上均优于传统 BP 算法。Khademi 等人[47]提出了一种新型糖尿病诊断系统，结合了 SVM、KNN 和鲸鱼优化算法（Whale Optimization Algorithm，WOA）。WOA 用于为分类器生成权重，从而提高分类准确率。实证结果显示，该系统的准确率为 83%，比最好的现有分类器提高了 5%，并且相较于粒子群优化（Particle Swarm Optimization，PSO）提高了约 1%。Mishra 等人[48]提出了增强自适应遗传算法（EAGA），用于优化糖尿病诊断的数据集，并与 MLP 结合使用。该方法在多个数据集上表现优异，最高准确率达到 97.76%。Kamel 等人[49]提出了一种基于蚂蚱优化算法（Grasshopper Optimization Algorithm，GOA）的特征选择方法，以提高糖尿病诊断的准确性。他们在 PIMA Indian 数据集上应用了这一方法，并使用 SVM 算法达到了 97%的准确率。Jovanovic 等人[50]提出了一种基于 XGBoost 算法的糖尿病分类新方法，利用行星优化算法（Planetary Optimization Algorithm，POA）来优化 XGBoost 的超参数，以提高分类性能。他们的研究表明，这种方法在分类任务中的效果优于其他现有算法。Navazi 等人[51]提出了一种混合算法，通过 PSO 进行特征选择，并使用遗传算法（Genetic Algorithm，GA）优化 SVM 的超参数。这种方法在糖尿病早期诊断中表现优异，准确率达到 93%。Jibril 等人[52]采用 PSO 优化 SVM 和 LR 的超参数，以提高糖尿病诊断的准确性，使用 PIMA Indian 糖尿病数据集进行实验，优化后的 SVM 模型达到 98.67%的准确率，LR 模型达到 97%的准确率，显示了该方法在糖尿病诊断中的有效性。Arsyadani 等人[53]使用 KNN 对糖尿病进行诊断。通过两阶段变异灰狼优化算法（TMGWO）选择最佳特征子集，并采用合成少数类过采样技术对数据进行平衡处理。通过 10 折交叉验证，方法在糖尿病分类中的准确率达到 98.85%。

从上述讨论可以看出，群智能优化算法通常用于模型超参数优化和特征选择。与传统方法相比，它们具有更强的全局搜索能力和适应性，能够有效避免局部最优。同时，这些算法在处理高维数据时能够减少特征维度，提升模型准确性。此外，群智能优化算法与集成学习的结合，也被证明是一种有效提高糖尿病诊断性能的方法。

通过分析国内外的研究现状，可以得出以下结论：

（1）糖尿病的诊断仍然面临挑战，尤其是在准确性和一致性方面。这导致许多患者未能在早期得到适当治疗，影响了健康管理效果。

（2）有效的特征选择能够简化数据处理过程，减少噪声和冗余信息。糖尿病诊断所涉及的数据通常具有高维度，很多特征可能是冗余的或不相关的。通过筛选出与糖尿病诊断最相关的特征，可以有效提升模型的性能。

（3）超参数选择对模型效果有显著影响。超参数配置直接影响机器学习模型的性能。合适的超参数可以提高模型的准确度和鲁棒性，而不当的选择则可能导致性能下降。因此，如何选择最优的超参数组合，是一个值得深入研究的问题。

1.3 研究内容

糖尿病的诊断受多种因素的影响，诊断过程复杂且多变。随着现代医学信息化进程的推进，传统的人工诊断方法面临着数据处理量大、结果依赖经验等问题。因此，借助机器学习与人工智能算法来提升糖尿病诊断的准确性和效率成为了当前的研究趋势。针对现有糖尿病诊断方法的不足，本文通过特征选择、超参数优化及分类模型的提升，进行糖尿病诊断的研究，旨在构建更高效、更准确的诊断模型，并为临床医疗决策提供科学的支持。具体研究内容如下：

（1）基于改进沙猫群优化算法的糖尿病特征选择方法研究。首先，介绍了糖尿病诊断所使用的数据集，包括数据来源、特征描述及其统计特性，随后，在原始沙猫群优化算法的基础上进行改进，以增强其全局搜索能力和收敛性能，并通过基准函数测试验证改进算法的有效性。接着，利用改进的沙猫群优化算法对高维糖尿病数据进行特征选择，筛选出与糖尿病诊断最相关的特征，从而减少冗余信息和噪声，提升模型的可解释性和泛化能力。最后，将筛选得到的特征集与传统特征选择方法进行对比实验，评估不同方法的性能，以验证所提方法的有效性和优势。

（2）基于优化的 Stacking 集成学习的糖尿病诊断算法研究。先介绍了基于分类错误率的权重策略对 Stacking 集成学习进行改进，其次对原始的麻雀优化算法进行改进并加以算法验证，接着，经过斯皮尔曼相关系数分析不同机器学习模型之间的关联程度，并选取了 LR、DT、KNN 和 XGBoost 作为 Stacking 集成学习的基学习器，经过实验分析，最终选定了 SVM 作为 Stacking 模型的元学习器，并基于优化的 Stacking 集成学习方法构建糖尿病诊断模型。此外，为了进一步提升 Stacking 集成学习模型的性能，采用了改进的麻雀优化算法对这些学习器的超参数进行了优化，以验证所提模型的优势。

（3）糖尿病诊断系统的设计与实现。该系统旨在构建一个高效、直观的糖尿病诊断平台，辅助医疗人员进行疾病诊断，并提升患者的健康管理能力。后端采用 Django 框架处理数据存储与模型调用，前端基于 Vue 构建交互界面，确保良好的用户体验。系统集成了基于优化的 Stacking 集成学习的诊断模型，并结合优化算法提升诊断准确性。支持患者体检数据管理、智能诊断和结果可视化，提供安全稳定的服务，助力精准医疗。

1.4 技术路线

本文聚焦于糖尿病诊断研究，主要从特征选择和分类模型优化两个方面展开，以提高诊断的准确性和稳定性，首先，针对糖尿病数据的特征冗余问题，采用改进的沙猫群优化算法进行特征筛选，从而提升模型的有效性并降低计算复杂度。随后，构建基于分类错误率的权重策略的 Stacking 集成学习分类模型，通过集成多种基学习器，提高模型