基于多模型的胎儿染色体异常判定的探究

魏婧，刘昱贤，皮哲豪，李乃医 *
广东海洋大学 数学与计算机学院，广东 湛江 524000
DOI:10.61369/ASDS.2026010013

摘　　要： 本研究针对女胎染色体异常判定中模型易受性别因素干扰、泛化性能不足的问题，构建多模型分析框架。首先对比LDA、QDA、GNB模型的决策边界判定效果。然后分析DNN、改进型MLP及融合注意力机制的Att–MLP模型的阈值判定结果。最后创新性引入神经网络迁移学习方法，将非线性注意力机制与迁移学习结合。研究表明，LDA与Att–MLP模型准确率均超90%，且Att–MLP模型特异度达97.25%，有效避免了误判问题。迁移学习虽效果有限，但验证了性别间异常特征的可迁移潜力，为后续研究提供了方向。

关 键 词： 线性判别分析；全连接神经网络；改进多层感知机；注意力机制；迁移学习

Research on the Determination of Fetal Chromosomal Abnormalities Based on Multiple Models

Wei Jing, Liu Yuxian, Pi Zhehao, Li Naiyi*
School of Mathematics and Computer Science, Guangdong Ocean University, Zhanjiang, Guangdong 524000

Abstract: This study addresses the issues of gender factor interference and insufficient generalization performance in the determination of chromosomal abnormalities in female fetuses by constructing a multi–model analysis framework. Firstly, the decision boundary determination effects of LDA, QDA, and GNB models are compared. Then, the threshold determination results of DNN, improved MLP, and the attention mechanism integrated Att–MLP model are analyzed. Finally, the neural network transfer learning method is innovatively introduced, combining the nonlinear attention mechanism with transfer learning. The research shows that the accuracy rates of both LDA and Att–MLP models exceed 90%, and the specificity of the Att–MLP model reaches 97.25%, effectively avoiding misjudgment problems. Although the effect of transfer learning is limited, it verifies the transfer potential of abnormal features between genders, providing a direction for subsequent research.

Keywords: linear discriminant analysis; fully connected neural network; improved multi–layer perceptron; attention mechanism; transfer learning

引言

目前，无创产前检测（NIPT）已广泛应用于胎儿染色体异常早期筛选方面[1–3]。然而，女胎染色体异常的判定因异常样本稀缺、特征维度与男胎存在显著差异等原因，仍面临数据不足、判定精度欠佳等问题。传统统计方法常依赖线性假设，难以捕捉女胎样本中相关指标间复杂的非线性特征关系。近年来，机器学习技术在医学领域展现出强大潜力[4]，注意力机制的融合使模型能够聚焦关键特征。然而，将线性判别分析与迁移学习结合的研究尚不多见[5]，且女胎异常判定问题缺乏系统性方法探究。

为此，本文针对女胎异常样本稀缺的痛点，通过多模型比较、融合注意力机制的新模型构建、迁移学习探索及提出“初筛 + 复核”分级决策策略，为女胎异常判定提供一种高精度、高稳健性的方法，并尝试突破女胎数据稀缺对检测性能的制约问题。

基金项目：2022年度国家自然科学基金项目“基于广义相依删失数据小波估计的构建与统计推断”（12161075）；2024年广东省自然科学基金项目“排序集抽样下相依删失数据经验似然推断”（2024A1515011258）。
作者简介：
魏婧，广东海洋大学数学与计算机学院，本科生；
刘昱贤，广东海洋大学数学与计算机学院，本科生；
皮哲豪，广东海洋大学数学与计算机学院，本科生。
通讯作者：李乃医，广东海洋大学数学与计算机学院，博士，教授，硕士生导师，研究方向为数理统计及其应用。

一、相关研究

以往关于女胎染色体异常的判定研究，多集中于判定模型的开发，强调了染色体异常判定对影响染色体疾病的筛查与早期诊断的重要性。早期研究通过传统线性判别模型进行判定，但该模型往往难以捕捉多维数据之间存在的复杂非线性关系。

线性判别分析模型与神经网络迁移学习的结合是一项重要突破，可充分发挥线性判别分析模型的强可解释性与神经网络捕捉复杂特征的能力，通过跨性别数据的迁移学习方法，有效弥补女胎数据较少、样本特征学习不充分等不足，从而提高预测精度。

二、数据预处理

本研究的实证基础是2025年全国大学生数学建模竞赛公开数据集。该数据集包含的男女胎数据样本均有27个不同的特征属性，本文系统地进行了数据预处理工作。

缺失值处理：孕妇 BMI 有缺失情况，通过身高与体重补全缺失的 BMI 数值。

异常值处理：为防止测试质量出现问题，删除 GC 含量远低于40% 的数据。

数据增强采用“基于噪声添加的自定义数据增强 +RandomOverSampler 过采样处理 + 样本权重调整”的组合策略：少数类增强倍数设为1，过采样处理增加少数类样本数量，初步平衡类别分布；自动平衡样本权重强化模型对异常类样本的关注。

数据划分采用模型差异化策略：传统判别模型按8:2比例随机分层划分训练集与测试集。神经网络相关模型按7:2:1比例随机分层划分训练集、测试集与验证集。

三、建模与求解

（一）传统判别模型

本文通过比较线性判别分析、二次判别分析、高斯朴素贝叶斯三种模型的性能指标（如表1）[6]，发现 LDA 模型表现更优，选择该模型作为后续分析的基准模型。

表1：不同模型对比表

模型	交叉验证平均得分	测试集准确率
线性判别分析（LDA）	0.9120	0.9091
二次判别分析（QDA）	0.9021	0.8843
高斯朴素贝叶斯（GNB）	0.8821	0.8926

由表2可知，LDA 模型训练集和测试集的准确率差距极小，均在0.9以上，说明模型泛化能力稳定，且训练集和测试集的 AUC 值均高于0.8，说明模型分类效果好。

表2：线性判别分析评估指标

评估指标	训练集	测试集
准确率	0.9170	0.9091
精确率	0.7895	0.8333
F1分数	0.4286	0.4762
AUC-ROC	0.8113	0.8277

为验证模型的正确性，图1和图2分别从可视化和量化角度展示了 LDA 模型的判定表现。由图1、图2可知模型预测准确率为0.9154，特异度为532/537=99.07%，说明模型整体预测效果较好。

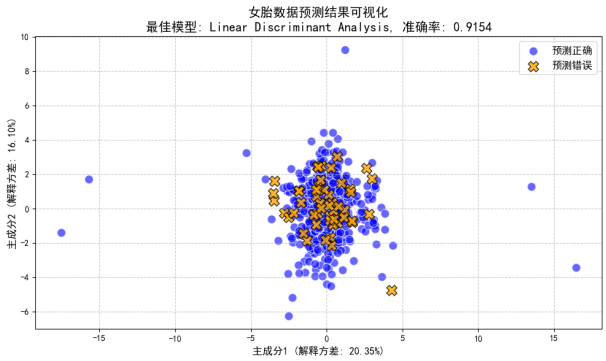

图1：女胎数据预测结果可视化

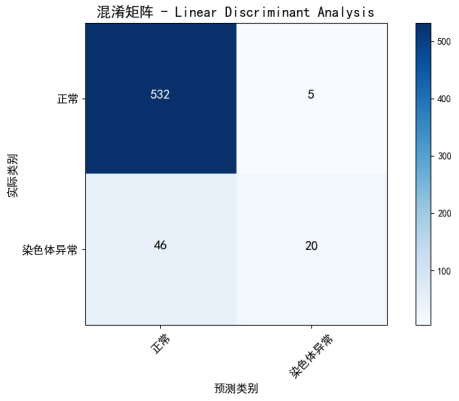

图2：女胎数据预测结果的混淆矩阵热图

（二）神经网络模型

针对染色体异常的判定问题，本文还建立了神经网络模型，分别为全连接神经网络模型、针对性改进多层感知机与融合注意力机制的针对性改进多层感知机。

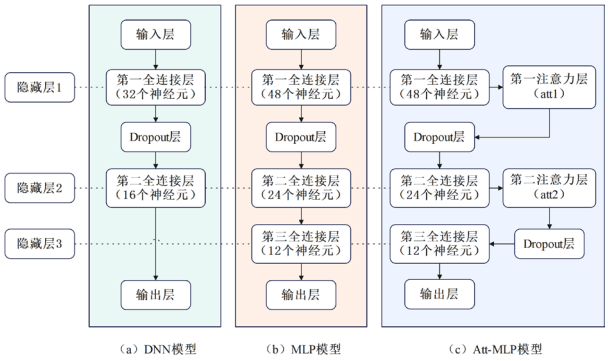

图3：三种神经网络模型结构对比图

1. 全连接神经网络（DNN）模型

采用 Sequential 线性堆叠结构，包含输入层、2个全连接层（均使用 ReLU 激活函数）、1个 Dropout 层与输出层。该模型结合 L2正则化抑制权重过大、Droupout 层随机失活20% 节点避免过拟合。DNN 模型阈值为0.3680。

表3：全连接神经网络（DNN）模型评估指标

	准确率	灵敏度	特异度	F1分数
DNN 模型	0.9007	0.6250	0.9817	0.7407

2. 针对性改进多层感知机（MLP）

针对DNN模型的不足，MLP模型采用Sequential顺序结构，包含输入层、3个全连接层、1个Dropout层与输出层。全连接层激活函数由ReLU替换为GeLu，以解决ReLU激活函数的“神经元死亡”问题与“梯度消失”风险。MLP模型阈值为0.3911。

表4：针对性改进多层感知机（MLP）评估指标

	准确率	灵敏度	特异度	F1分数
MLP模型	0.9078	0.6562	0.9817	0.7636

3. 融合注意力机制的针对性改进多层感知机（Att-MLP）

注意力机制源于生物视觉系统，在视觉上减少对冗余信息的关注，以聚焦更多的关键信息。通过分配特征权重将关键数据分配更多的权重，以获取更多有效信息[7]。

针对MLP模型的不足，Att-MLP模型采用“全连接层+残差式注意力”的组合架构，包含输入层、3个全连接层、2个残差注意力层、2个Dropout层与输出层。

该模型核心点在于双层注意力机制。att1、att2分别作用于浅层原始特征与深层抽象特征。“特征重要性评估+残差特征增强”的残差式设计，模拟了临床医生先关注核心病理指标、后综合多指标异常关联规律的诊断逻辑，提升了对染色体异常判定的准确性与临床适配性。Att-MLP模型阈值为0.3478。

表5：融合注意力机制的针对性改进多层感知机（Att-MLP）评估指标

	准确率	灵敏度	特异度	F1分数
Att-MLP模型	0.9128	0.7500	0.9725	0.8219

表6：Att-MLP特征权重分配表

指标	特征权重
13号染色体的Z值	0.0162
18号染色体的Z值	0.0176
21号染色体的Z值	0.0176
X染色体的Z值	0.0213
……	……
13号染色体的GC含量	0.0166
18号染色体的GC含量	0.0236
21号染色体的GC含量	0.0182
X染色体浓度	0.0220

对比三大模型评估指标结果（表3、表4、表5），DNN模型识别异常样本能力较弱；MLP模型较DNN模型能更有效地捕捉异常特征，灵敏度有所提高；Att-MLP模型引入注意力机制，使模型自动聚焦关键特征，灵敏度显著高于DNN、MLP模型。

其中，Att-MLP模型的准确率达0.9128，特异度为0.9725，表明该模型在识别正常样本方面具有高可靠性，可有效减少怀正常胎儿的孕妇的检查流程。且该模型为多个关键特征赋予了合理权重，既体现了模型权重的分配符合医学中重点关注的染色体非整倍体类型，又验证了模型构建的科学性与合理性。

（三）神经网络模型的迁移学习方法

目前，我们已经构建了LDA、Att-MLP两大模型，均展现出优异的分类效能，考虑到“染色体非整倍体”的异常特征可能存在“跨性别共性”，且女胎数据少，本文提出采用迁移学习方法，即利用男胎异常判定知识提升女胎异常判定性能的方法。

为验证“跨性别共性”的合理性与“迁移学习方法”的可行性，本文将LDA模型运用在男胎数据上，将预测值与实际值进行比较，同样从可视化和量化角度展示了LDA模型的判定表现。由图4、图5可知，该模型预测准确率为0.9122，特异度为948/956=99.16%，说明LDA模型在男胎数据上预测效果同样较好。

结果表明“跨性别共性”具有一定的合理性，所以选择使用结合了双模型优势的神经网络迁移学习方法。迁移学习精简了上述Att-MLP模型，通过对模型结构的优化来提高模型的泛化能力和运行效率，从而更适合处理目标数据少的小样本二分类。

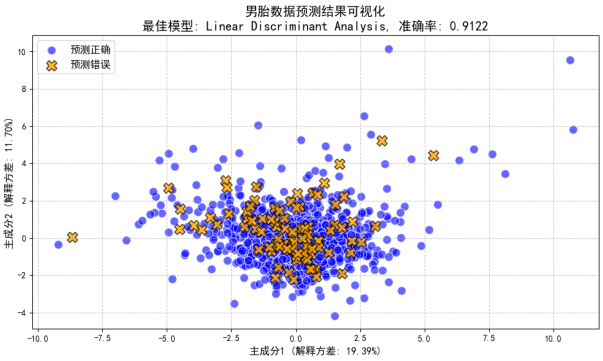

图4：男胎数据预测结果

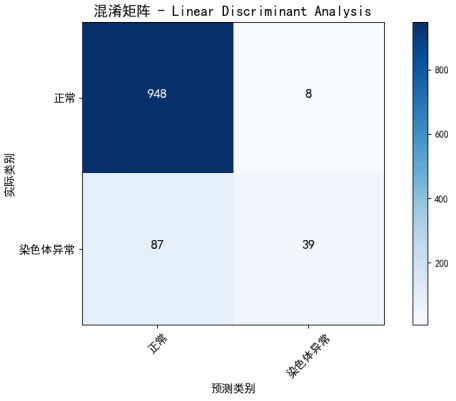

图5：男胎数据预测结果的混淆矩阵热图

本文设计先利用男胎数据训练模型。

表7：迁移学习特征权重分配表

指标	特征权重
13号染色体的Z值	0.0416
18号染色体的Z值	0.0418
21号染色体的Z值	0.0307
X染色体的Z值	0.0226
……	……
13号染色体的GC含量	0.0259
18号染色体的GC含量	0.0264
21号染色体的GC含量	0.0380
X染色体浓度	0.0278

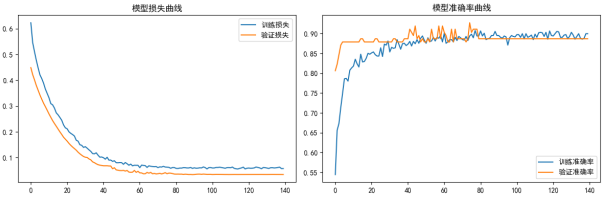

图6：染色体异常检测模型训练损失与准确率曲线

表7中，该模型对核心染色体（13号、18号、21号）的Z值等与染色体异常密切相关的特征赋予较高权重，符合医疗逻辑。图6是模型在训练阶段的监控指标，用于评估模型在男胎数据上的学习效果。模型在整个训练过程中表现出良好的收敛特性。且验证集曲线始终与训练集保持相似趋势，表明模型学习过程稳定，泛化能力可靠。

综上所述，基于男胎数据进行训练的迁移学习确实学到了染色体异常的核心特征等可复用的有效知识。本研究针对迁移学习模型的核心参数进行了针对性微调后[8]再将该模型运用于女胎数据实现异常判定，以提升预训练模型在女胎数据上的适应性。

由图7可直观看出模型预测特异度为339/537=63.13%，灵敏度为56.06%，漏诊率较高，风险较大。对比LDA模型，迁移学习准确率也较低，仅62.35%。

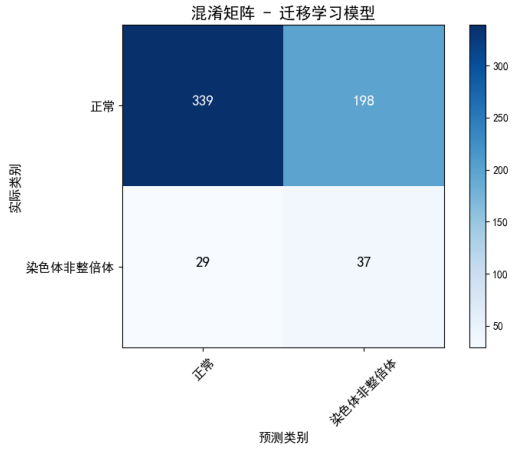

图7：迁移学习女胎数据预测结果的混淆矩阵

针对该问题，本文尝试通过优化迁移策略提升模型性能，采用渐进式两阶段迁移学习策略，第一阶段冻结特征提取器，仅对分类器层进行重新训练，第二阶段解冻特征提取器，实现目标域特征的适配。结果表明，模型准确率提升至71.48%，特异度达77.65%，但灵敏度下降至21.21%，存在显著类别偏置性。核心问题与具体改进方向为：

数据分布差异：男胎与女胎的染色体特征仍存在本质差异，迁移学习模型未能完全适配女胎数据特性。可引入领域自适应网络，通过对抗训练[9]实现特征空间分布对齐，增强模型跨性别泛化能力。

迁移学习策略不足：未对模型进行完善的训练结构优化与参数微调，导致模型在小数据集场景下收敛速度慢。可采用渐进式分层微调策略，通过分层解冻参数，使模型学习到目标域的特异性特征。

样本类别不平衡：女胎数据中异常样本仅占10.95%，类别失衡问题突出，导致模型无法聚焦于异常样本的特征提取。可采用生成对抗网络，使用生成合理的异常样本，再通过判别器区分真实数据与生成数据。

四、结论与讨论

本文通过多模型探究，发现LDA、Att-MLP模型体现了模型特性的差异，McNemar检验p值为0.0023，远小于显著性水平α=0.05，说明两模型间的性能差异显著。

表8：LDA与Att-MLP模型性能与统计检验结果

模型	准确率	灵敏度	特异度	检验统计量	p值
LDA	0.9154	0.3030	0.9907	9.3333	0.0023
Att-MLP	0.9128	0.7500	0.9725		

数据显示，LDA模型特异度高、可解释性强，适用于高特异度需求、误判容忍度低的临床初筛场景；Att-MLP模型灵敏度与综合性能更优，适用于高灵敏度需求的复核场景。针对构建胎儿异常判定的临床辅助系统，本文提出“初筛+复核”的分级决策策略：通过LDA模型快速扫描样本实现初筛，再通过Att-MLP模型处理可疑样本。该策略需应对计算复杂度相对较高、难以快速获得临床医生的信任与采纳等核心问题。

因此，我们提出以下建议：在模型架构方面，可引入轻量型Transformer变体[10]，降低临床设备的算力需求。在数据增强方面，可采用SMOTE-ENN混合采样法，实现合成样本的生成、噪声和冗余样本的剔除，以平衡类别分布。在验证体系拓展方面，可收集不同医院数据，验证“初筛+复核”策略的稳定性与模型的适配性。

为应对女胎数据稀缺所带来的建模挑战，本研究首次将非线性注意力机制与迁移学习结合，未来应持续探索迁移学习等跨性别共性挖掘方法。

参考文献

[1] 蒋丽雅，卢劭侃，杜佳恩，等.无创产前检测技术的发展与应用[J].临床医学研究与实践，2025,10(23):191-194.DOI:10.19347/j.cnki.2096-1413.202523047.

[2] 于丹丹，李奉瑾，姚欣雨，等.NIPT在不同慎用人群胎儿常见染色体非整倍体异常检测中的效能分析[J].中国计划生育和妇产科，2025,17(03):72-76+82.

[3] 鞠爱萍，孟祥荣，覃燕龄，等.无创产前检测在筛查胎儿染色体拷贝数变异中的应用价值[J].实用心电与临床诊疗，2025,34(05):665-671.DOI:10.13308/j.issn.2097-5716.2025.05.008.

[4]Alonso E，Beristain A，Burgos J，et al.Comparison of Machine Learning Algorithms to Predict Down Syndrome During the Screening of the First Trimester of Pregnancy[J].Applied Sciences (2076-3417)，2025，15(10).DOI:10.3390/app15105401.

[5] 车志勇.基于线性判别分析（FDA）的迁移学习方法[D].广东工业大学，2019.DOI:10.27029/d.cnki.ggdgu.2019.000284.

[6] 林宏赡.基于朴素贝叶斯、线性判别、二次判别分类算法的选股实证研究[D].山东大学，2018.

[7] 赵文海.基于自注意力机制和互通双流MLP的点击率预估模型研究[D].电子科技大学，2025.DOI:10.27005/d.cnki.gdzku.2025.004745.

[8] 王柯惟.基于深度学习的染色体异常检测研究[D].中南大学，2023.DOI:10.27661/d.cnki.gzhnu.2023.001790.

[9] 雷雨佳.基于卷积神经网络的染色体异常检测[D].湖南师范大学，2020.DOI:10.27137/d.cnki.ghusu.2020.001208.

[10] 赵秋博.基于Transformer特征关联融合小目标检测算法研究[D].西安电子科技大学，2024.DOI:10.27389/d.cnki.gxadu.2024.002573.

