Modeling and Simulation 建模与仿真, 2023, 12(2), 1321-1330
Published Online March 2023 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2023.122124

Hans汉斯

基于机器学习的阿尔兹海默症智能分类研究

郑　耀，杨亭亭，周志鹏

福建江夏学院电子信息科学学院，福建　福州

收稿日期：2023年1月28日；录用日期：2023年3月14日；发布日期：2023年3月21日

摘　要

阿尔兹海默症(AD)是一种使人记忆力衰退、大脑功能缓慢且逐渐变异的智力丧失表现的疾病。而早期诊断和预测AD可以延缓患者的发病，可能延长患者的寿命，对整个社会具有重要的科学意义。本文通过机器学习等方法对AD数据进行分析和建模，增添大脑特征指标，进行数据预处理，利用成对分类法的SVM实现AD的智能诊断，通过SVM得到的五分类的准确度达到0.71，高于随机森林的0.69与XGBoost提升树的0.706。并通过绘制ROC曲线，对于判断CN有64%准确率，判断AD有80%准确率，判断LMCI有68%准确率，判断SMC有58%准确率，判断EMCI有76%准确率，具有较好的测试效果。提取和发现影响AD诊断和预测的关键症状，并对患者类别进行分类，能够有效、快速地得到早期诊断和预测AD的患病概率，以达到预防AD的有效手段。基于成对分类法的SVM实现AD的智能诊断，并且能够用推广于更加多分类的问题。

关键词

随机森林，支持向量机，阿尔兹海默，智能分类

Intelligent Classification of Alzheimer's Disease Based on Machine Learning

Yao Zheng, Tingting Yang, Zhipeng Zhou

College of Electronics and Information Science, Fujian Jiangxia University, Fuzhou Fujian

Received: Jan. 28th, 2023; accepted: Mar. 14th, 2023; published: Mar. 21st, 2023

Abstract

Alzheimer's disease (AD) is a disease that causes memory loss and slow and gradual variation in brain function as a manifestation of intellectual loss. Early diagnosis and prediction of AD can delay the onset of the disease, potentially prolong the life span of patients, and is of great scientific importance to society as a whole. In this paper, we analyzed and modeled AD data by machine learning, added brain feature indicators, preprocessed data, and used SVM with pairwise classification method to achieve intelligent diagnosis of AD. The accuracy of the five classifications obtained by SVM

文章引用: 郑耀, 杨亭亭, 周志鹏. 基于机器学习的阿尔兹海默症智能分类研究[J]. 建模与仿真, 2023, 12(2): 1321-1330.
DOI: 10.12677/mos.2023.122124

reaches 0.71, which is higher than 0.69 of random forest and 0.706 of XGBoost boosted tree. The accuracy of the ROC curve is 64% for CN, 80% for AD, 68% for LMCI, 58% for SMC, and 76% for EMCI, which is a good test result. By extracting and discovering the key symptoms affecting the diagnosis and prediction of AD, and classifying the patient categories, we can effectively and quickly obtain early diagnosis and predict the probability of AD, so as to achieve an effective means of preventing AD. The SVM based on pairwise classification method enables intelligent diagnosis of AD and can be extended to more multi-categorical problems.

Keywords

Random Forest, Support Vector Machines, Alzheimer's Disease, Intelligent Classification

1. 引言

设 AD 是一种使人记忆力衰退、大脑功能缓慢且逐渐变异的智力丧失表现的疾病。该病严重可以直接影响患者认识和感觉心理综合功能的严重缺失，尤其严重的后果是患者大脑神经皮质生理机能的严重丧失，包括正常人的逻辑记忆力、判断能力、抽象思维表达能力等。AD 的特点是神经原纤维缠结、神经元和突触丢失、β 淀粉样斑块的累积，所导致脑功能和行为出现变化。早期诊断和预测 AD 可以延缓患者的发病，可能延长患者的寿命，对整个社会具有重要的科学意义。因此，有效的早期诊断和预测 AD 的患病概率是预防 AD 的有效手段[1]。通过人工智能计算方法对 AD 数据进行分析和建模，提取和发现影响 AD 诊断和预测的关键症状。由于数据集中不同的量表有各自的优势和局限性，因此需要综合的考虑各种表格，从而诊断和预测 AD。特征提取算法有的仅支持离散属性的特征变量、有的仅支持连续属性的特征变量、也有的两者皆可支持。本文根据有关特征对阿尔茨海默病进行智能化诊断并建立相关数学模型。用附加的大脑结构特征和认知行为特征来设计阿尔茨海默病的智能诊断，本文采用的是基于成对分类法的 SVM 实现 AD 的智能诊断，从简单的 SVM 二分问题向高维的五分类问题逐渐优化，使得问题简单化然后对五类数据分别探索并验证模型的效果。

2. 模型假设与符号说明

1) 假设本题附件中提供的数据均真实可信。
2) 假设每个样本间相互独立、互不关联，并且他们的各指标之间互不干扰(表 1)。

Table 1. Symbol description
表 1. 符号说明

符号	说明	单位
R	相关系数矩阵	/
r_s	斯皮尔曼相关性系数	/
$K(x_i, x_j)$	SVM 核函数	/
J	松弛变量	/
β_i	回归系数	/
A	健康状况	/

3. 基于成对分类法的 SVM 实现 AD 的智能诊断

3.1. 模型原理介绍以及模型的优点

3.1.1. 模型原理介绍

支持向量机建立于结构风险最小原理的基础至上，以间隔最大化为学习策略，其算法本质是求解凸二次规划的最优化算法。由于支持向量机依靠接近平面的若干支持向量建立分界线或超平面，因此能够凭借有限的样本信息获取现有条件下的全局最优解，避免了神经网络方法可能面临的样本过少的局部最优解问题，同时具有较好的泛化能力。如下图所示，即为简单的二分类支持向量机。

3.1.2. SVM 的优势

SVM 的最终决策函数只由少数的支持向量所确定，计算的复杂性取决于支持向量的数目，而不是样本空间的维数，这在某种意义上避免了“维数灾难”。其次 SVM 主要擅长应付样本数据线性不可分的情况，主要通过核函数和松弛变量来实现。其次由于分类器仅由支持向量决定，SVM 还能够有效避免过拟合。

3.1.3. 敏感性分析

在所有候选的参数选择中，通过循环遍历，尝试每一种可能性，表现最好的参数就是最终的结果。先将数据集划分为 10 份，每次取其中 9 份用来训练，1 份用于测试，最终得到十次测试结果的平均值(外层交叉验证)。每次在用 9 分数据进行训练时，如果模型需要调参，则可进一步在训练过程当中进行交叉验证(内层)。通过每次取 9 份中的 8 份数据用于训练，剩余的一份用于验证，用 9 次的均值来对当前参数做出评估(图 1)。

结合数据集观察每个数据特征在随机森林中的每颗决策树上做了多少贡献，然后取平均值，最后对比特征之间的贡献大小。总结一下就是：特征重要性是指，在全部单颗树上此特征重要性的一个平均值，而单颗树上特征重要性计算方法事：根据该特征进行分裂后平方损失的减少量的求和。

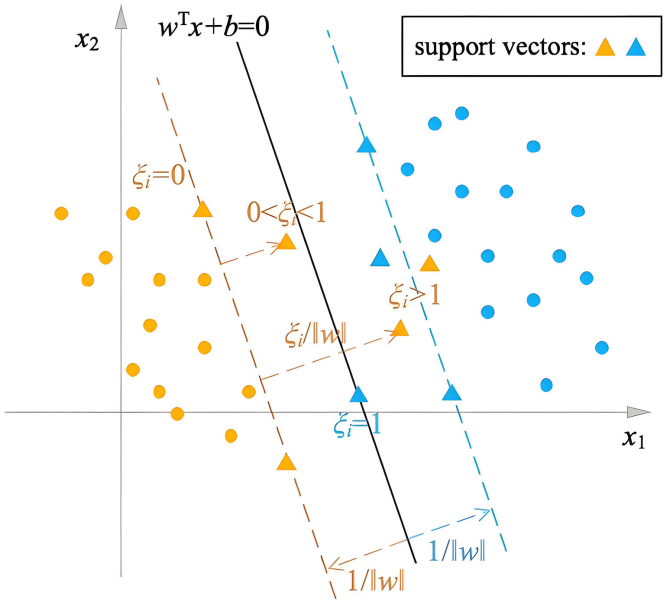

Figure 1. Binary classification support vector machine
图 1. 二元分类支持向量机

特征 x_j 在整个模型中的重要程度为：

$$\hat{J}_j^2 = \frac{1}{M}\sum_{m=1}^{M}\hat{J}_2^2(T_m)$$

其中，M 是模型中树的数量。特征 x_j 在单独一个树上的特征重要度为：

$$\hat{J}_j^2(T) = \sum_{t=1}^{L-1}\hat{I}_t^2 \ (v_t = j)$$

3.2. 数据预处理

选取了：

'CDRSB'，'ADAS13'，'MMSE'，'RAVLT_forgetting'，'RAVLT_perc_forgetting'，'LDELTOTAL'，'FAQ'，'mPACCdigit'，'mPACCtrailsB'，'Ventricles'，'Hippocampus'，'Entorhinal'，'Fusiform'，'MidTemp'，'ICV'

十六个指标[2]作为自变量，DX_bl 作为因变量。即指标的数量是 2425，维度为 16。

Step1：数据的归一化[3]

$$X_{ncw} = \frac{x_i - u_i}{\partial_i}$$

其中 x_i 是第 i 个指标的权重，u_i 是第 i 个指标的平均值，∂_i 为第 i 个指标的标准差。

Step2：使用网格搜索法找到最优内核函数

支持向量机核函数中，不同的超参数设置，机器学习给出的结果也会不一样，因此，也会影响到对结果的评价指标。而人们往往会追求一个“最好”的结果，因此，就需要在众多超参数的取值范围中选取一个“最优”的值进行设置。因此，使用网格搜索等手段，均是为了寻找好的超参数。利用网格搜索技术调整各项参数以取得最好的评价分数。

支持向量机中的内核函数有：

线性核函数：

$$K(x_i, x_j) = x_i^{\mathrm{T}} x_j$$

多项式核函数：

$$K(x_i, x_j) = \left(\gamma x_i^{\mathrm{T}} x_j + b\right)^d$$

高斯核函数：

$$K(x_i, x_j) = e\left(-\gamma \left\| x_i - x_j \right\|\right)$$

Sigmoid 核函数：

$$K(x_i, x_j) = \tanh\left(\gamma x_i^{\mathrm{T}} x_j + b\right)$$

Step3：建立五分类的二次型[4]规划问题：

目标函数：$f\left(x, w(\ )\operatorname{sgn}\sum_{i=1}^{N} W_i\left(K(x, x_i) + b\right)\right)$；

优化目标函数为：$J = w^{\mathrm{T}} w = \|w\|^2$；

约束条件：$s.t. y_i = \left[\sum_{i=1}^{N} w_i k(x_j, x_i) + b\right] \geq 1,\ j = 1, 2, 3, \cdots N$。

在目标函数中，为了提升准确度，消除噪声与异常样本的影响，引入松弛变量，表达式如下：

$$J = \frac{1}{2} w^T w + C\sum_{i=1}^{N}\zeta_i$$

$$\begin{cases} y_i\left[\sum_{i=1}^{N} w_i k\left(x_j, x_i\right)+b\right] \geq 1-\zeta_i \\ i=1,2,\cdots,N \\ \zeta_i \geq 0 \end{cases}$$

为更容易解决该目标优化问题，本文采用引入拉格朗日乘子的方式，将优化问题转化为对偶问题其表达式如下：

$$\max \sum_{i=1}^{m} a_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j y_i y_j x_i^{\mathrm{T}} x_j$$

$$s.t. \sum_{i=1}^{m\Sigma} a_i y_j = 0,\ 0 \leq a_i \leq C,\ i=1,2,\cdots,m$$

Step4：绘制 ROC 曲线。

ROC 曲线正是通过不断移动分类器的“阈值”：

1) 假设已经得出一系列样本被划分为正类的概率 Score 值，按照大小排序。

2) 从高到低，依次将“Score”值作为阈值 threshold，当测试样本属于正样本的概率大于或等于这个 threshold 时，我们认为它为正样本，否则为负样本。

3) 每次选取一个不同的 threshold，得到一组 FPR 和 TPR，以 FPR 值为横坐标和 TPR 值为纵坐标，即 ROC 曲线上的一点。

本文是一个五分类问题，可以根据所分的类别和测试样本的数量(2425)的概率矩阵，从而计算各个阈值的假正例率(FPR)和真正例率(TPR)，从而绘制出 5 条 ROC 曲线。最后对 5 条 ROC 曲线取平均，即可得到最终的 ROC 曲线。

3.3. 模型的求解

3.3.1. 核函数的选择

将数据带入建立好的的模型，通过 python 求解，得出当核函数为 linear，惩罚系数 C = 1，核函数系数为 scale 时，F1-score 为 0.71，为最优结果，评估表如下：(表 2)。

Table 2. Forecast results
表 2. 预测结果

label	precision	recall	F1-score
0	0.53	0.9	0.67
1	0.93	0.75	0.83
2	0.83	0.9	0.86
3	0.64	0.75	0.69
4	0	0	0
accuracy			0.71
macro avg	0.59	0.66	0.61
weighted avg	0.63	0.71	0.65

3.3.2. ROC 曲线

一个特定的分类器[5]和测试数据集，每一个实例都会得到一个分类结果，通过统计，利用上述公式，可以得到一组 FPR 和 TPR 结果，通过 python 绘制出 ROC 曲线：(图 2)。

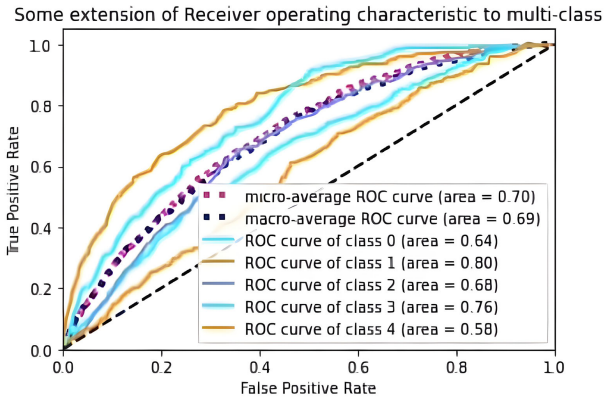

Figure 2. ROC curve
图 2. ROC 曲线

由上图中可以看出当核函数为 linear，惩罚系数 C = 1，核函数系数为 scale 时，本分类器对于判断 CN 有 64%准确率，判断 AD 有 80%准确率，判断 LMCI 有 68%准确率，判断 SMC 有 58%准确率，判断 EMCI 有 76%准确率，具有较好的测试效果。

3.4. 随机森林求解

我们将自变量作为输入，因变量作为目标导入随机森林工作箱，并按照一定的比例将这些样本分为两类进行处理：训练集(80%)、测试集(20%)。

3.4.1. 自变量和因变量的选择

其中自变量选取 AGE、PTGENDER、pteachat、PTRACCAT、MMSE、ADAS11、ADAS13、RAVLT_immediate、RAVLT_learning、RAVLT_forgetting、RAVLT_perc_forgetting、FAQ、mPACCdigit、ADASQ4_bl、LDELTOTAL_BL，共 15 个自变量；选择 DX_bl 作为因变量。

3.4.2. 随机森林决策树数量及模型训练方法的确定

接下来我们需要选定决策树的数量以及模型的训练方法来完成对模型的训练。经过大量的尝试，最终我们选用决策树数量为 71，模型训练方法为 ExtraTreesClassifier。

3.4.3. 随机森林模型的训练结果与模型评估

1) 各个参数的选择(表 3)

Table 3. Correlation of Alzheimer's disease with data characteristics
表 3. 阿尔茨海默病与数据特征的相关性

数据的特点	相关性得分
LDELTOTAL_BL	18.81%
mPACCdigit	14.30%
FAQ	8.53%
ADAS13	8.85%
MMSE	8.88%
AGE	6.98%
ADAS11	6.66%

Continued

RAVLT_immediate	5.40%
PTEDUCAT	5.35%
RAVLT_perc_forgetting	4.83%
RAVLT_learning	3.89%
ADASQ4_bl	3.88%
RAVLT_forgetting	3.49%
PTRACCAT	1.34%
PTGENDER	1.12%

2) 模型的评估

采用 sklearn 库中 metrics 的 classification_report 方法对模型准确率进行求解，经过大量的尝试得出随机森林的 n_estimators 选择为时 81，模型准确度最高。如图 1~3 所示：

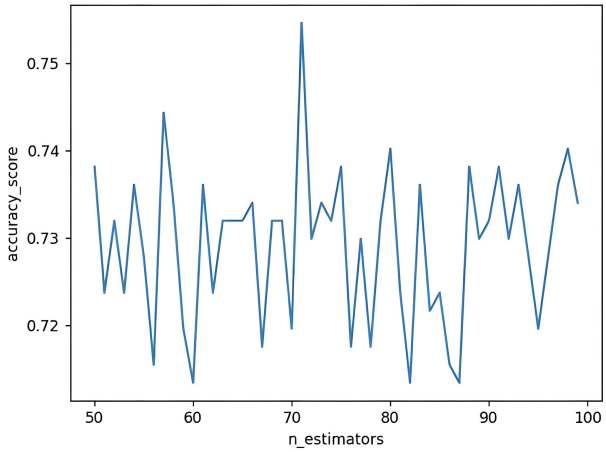

Figure 3. Model evaluation score
图 3. 模型评价评分

观察上图可得出整个模型震荡幅度较大，且在随机森林 n_estimators 选择为时 81，模型准确度最高。最终结果聚为五类和三类，可视化效果如图 4 所示。

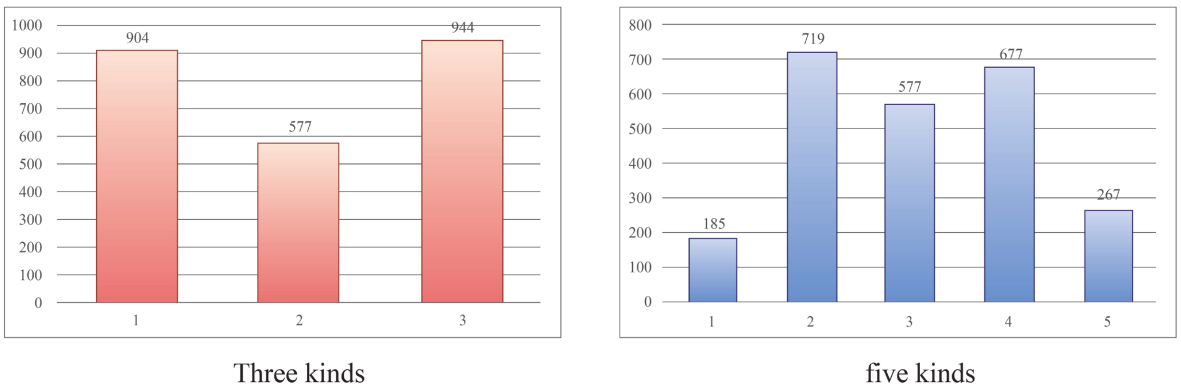

Figure 4. Population clustering chart
图 4. 人群聚类图

3.5. 多元神经网络[6]模型的求解

我们首先需要定量地研究不同数据特征对于阿尔兹海默症的影响，考虑引入人工神经元模型，图 5 展示了人工神经网络基本单元的神经模型。

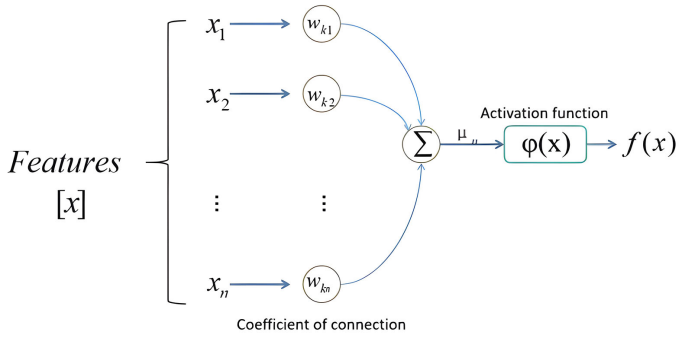

Figure 5. Neural network model
图 5. 神经网络模型

其中，$x_1, x_2, \cdots, x_n$ 属于样本特征，$w_{k1}, w_{k2}, \cdots, w_{kn}$ 是权重(也叫连接系数)，μ_n 是指各个样本特征的权重值之和，$\varphi(x)$ 时指激活函数这里我们采用默认内置 relu 函数：上述模型可以归纳成为以下公式：

$$\mu_n = \sum_j w_{kn} x_n,\ f(x) = \varphi(\mu_n),\ n = 1,2,\cdots,16$$

我们选取了 LDELTOTAL_BL，mPACCdigit，FAQ，ADAS13，AGE，MMSE，ADAS11，PTEDUCAT，RAVLT_perc_forgetting 共 9 项数据特征作为大脑结构特征和认知行为特征的特征标签；接下来我们需要选定隐藏神经元的个数以及模型的训练方法[7]来完成对模型的训练[8]。经过大量的尝试，最终我们选用隐藏神经元个数为 9，模型训练方法为贝叶斯正则化。阿尔茨海默病的智能诊断方案参考区间最终如表 4 所示。

当个人同时所检测出来的各项数据特征同时满足上表的参考范围可初步断定该个体患有阿尔兹海默症。

Table 4. Intelligent diagnosis solutions
表 4. 智能诊断方案

判断指标	最优参考范围
LDELTOTAL_BL	0.0~23.0
mPACCdigit	−23.69~6.30409
FAQ	0.0~30.0
ADAS13	0.0~54.68
AGE	50.4~91.4
MMSE	16.0~30.0
ADAS11	0.0~42.68
PTEDUCAT	4.0~20.0
RAVLT_perc_forgetting	−400~100

4. 关于阿尔兹海默的诊断标准

4.1. 关于 ApoE 基因方面的诊断

根据本文的研究并结合相关临床案例分析[9]，ApoE (Apolipoprotein E，载脂蛋白 E)基因是与 AD 关

系最相关的基因。ApoE 有 3 个等位基因：ApoE-Ɛ2、ApoE-Ɛ3 和 ApoE-Ɛ4。Farrer 等的研究发现，91%的 ApoE-Ɛ4 纯合基因携带者在 68 岁左右发病，约 47%的 ApoE-Ɛ4 杂合基因携带者在 76 岁时发病，而在 ApoE-Ɛ4 非携带者中，仅 20%在 85 岁时才发生 AD。研究结论是患者携带的 Ɛ4 基因的数量越多，患者的发病时间就越早。研究表明，ApoE 与 Aβ 代谢有密切关系，其中，ApoE-Ɛ4 对 Aβ 沉积所形成的老年斑有较强影响，并促进了脑淀粉样血管病的发生，两者是 AD 的最主要病理特征。

Table 5. Pathological characteristics of exercise affecting cognitive dysfunction in clinical experiments [3]
表 5. 临床实验中运动影响认知功能障碍患者的病理特征[3]

干预对象	运动强度	运动周期	样本量(例)	年龄	检测与评估方式	病理改变
携带 AD 风险基因的老年人	中等强度跑步	26 周	23	64.9 ± 5	VO2 峰、MMSE、CVLT、血样检验。	运动后患者心肺功能改善、认知功能改善
轻度 MCI 患者	中等强度舞蹈	12 周	36	72.9 ± 5.6	WMS-III、RBANS、BNT、MMSE、HADS、BBS。	运动后患者语言识别记忆、平衡能力和步态表现均改善，认知功能改善。
轻度 MCI 患者	中等强度快走上坡	12 个月	30	2657	磁共振、ADAS-Cog、威斯三、考特。	运动增加海马区血流量，改善患者记忆功能
轻度 MCI 患者	中高强度快走	12 个月	70	64 ± 6.6	VO2 峰、CBFV、CVLT-II、D-KEFS。	运动提高患者 VO2 峰，改善脑血流量，提高认知能力。
轻度 MCI 和 AD 患者	中等强度自行车	6 个月	87	79	MMSE、TMT、6MWT、ADAS-Cog、BMI、血样检验。	运动改善患者心血管功能，减缓认知能力下降。
轻度 AD 患者	有氧运动	4 年	376	80	ADAS-Cog、CDR-SOB	运动后患者认知功能、语言流畅性和情景记忆能力均得到改善。
AD 患者	高强度抗阻训练	6 个月	101	69.5 ± 6.6	ADAS-Cog、WAIS-III、MRI。	运动改善患者认知功能。
AD 患者	有氧运动	6 个月	96	77.4 ± 6.8	ADAS-齿轮。	运动减轻患者认知功能障碍。

4.2. 对患者进行量表诊断[9]

临床医师需要充分考虑患者的海马体的情况，还有受访时的年龄、是否患过精神疾病和 ADAS13，LDELTOTAL，LIMMTOTAL，MOCA，RAVLT.perc.forgetting，COPYSCOR，MMSE，EcogSPTotal，FAQ，EcogPtTotal 量表。并且对于患者的随访也要有针对性地选择这些量表测验。DBN 的结果表明，对临床医师判断患者患病时长的建议，要注意海马体和内嗅随着时间的变化情况，以及 ADAS13 量表等。这为临床医师通过已知的特征信息来辅助诊断患者实际所处的患病阶段有帮助[10]。

4.3. 关于阿尔兹海默症的早期干预

关于 MCI 与 AD 早期患者的干预措施主要通过运动进行影响认知功能障碍患者(表 5)。

参考文献

[1] 魏彩锋, 曾宪华. 基于字典对学习的轻度认知功能障碍识别[D]: [硕士学位论文]. 重庆: 重庆邮电大学, 2018.

[2] 罗佩琪, 康佳霞. 基于多特征融合的阿尔茨海默病早期诊断及预测方法研究[D]: [硕士学位论文]. 北京: 北京邮电大学, 2021.

[3] 李彩, 范炤. 基于机器学习的阿尔兹海默症分类预测[J]. 中国医学物理学杂志, 2020, 37(3): 379-384.

[4] 丘致榕, 丁雪梅, 杨洪钦. 不同卷积神经网络对稳定型与进展型轻度认知障碍识别的比较[D]: [硕士学位论文]. 福州: 福建师范大学, 2021.

[5] 胡喜园, 邓洪敏, 徐泽林, 尹双才, 朱麒麟. 基于改进残差网络的阿尔兹海默症分类[J]. 计算机应用, 2022, 42(SI): 71-79.

[6] 刘丽萍, 仲伟伟, 董文南, 张文英. 基于深度学习的阿尔兹海默症自动诊断方法研究[J]. 科学咨询, 2021(27): 54-55.

[7] 杨邦坤, 汪乐生, 聂颖, 熊文平. 基于机器学习的阿尔兹海默症初期行为辨识方法[J]. 生物医学工程研究, 2021, 40(2): 121-125. https://doi.org/10.19529/j.cnki.1672-6278.2021.02.03

[8] 楚阳, 徐文龙. 基于计算机辅助诊断技术的阿尔兹海默症早期分类研究综述[J]. 计算机工程与科学, 2022, 44(5): 879-893.

[9] 何慧萍, 何尧苇, 沈宗霖, 宋肖肖, 李葆罗, 姜红燕. 阿尔茨海默病与轻度认知功能障碍患者精神行为症状比较分析[J]. 昆明医科大学学报, 2022, 43(9): 19-23.

[10] 陈艳洁, 张源进, 吕忠宽, 李亚明. 阿尔茨海默氏病的防治干预[C]//中国中西医结合学会. 中国中西医结合学会第八届虚证与老年医学专业委员会、中国老年学和老年医学学会中西医结合分会、江苏省中医药学会老年医学专业委员会 2019 年学术年会论文集. 2019.