PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2021,48(8):888~897
www.pibb.ac.cn

基于遗传数据的机器学习在阿尔茨海默病研究中的应用*

金 宇[1,2] 姚旭峰[2,3]** 韩立婷[1,2] 赵从义[1,2] 黄 钢[2,3]
([1] 上海理工大学医疗器械与食品学院，上海 200082；[2] 上海健康医学院医学影像学院，上海 201308；
[3] 上海健康医学院，上海市分子影像学重点实验室，上海 201308)

摘要 阿尔茨海默病（Alzheimer disease，AD）是一种神经退行性疾病，其发病与遗传和环境因素相关，约70%由遗传因素引起，但其发病机制尚不清楚. 随着高通量测序技术的出现，利用机器学习（machine learning，ML）技术处理遗传数据的研究成为了当前热点. 本文综述ML在AD中的应用研究，主要包括：遗传数据与影像、临床、组学等多模数据结合的AD诊断和预后；对单核苷酸多态性（single nucleotide polymorphism，SNP）数据挖掘发现与AD风险相关基因的遗传变异分析；与AD发病机制密切相关的基因表达谱分析. 最后，应用高质量、综合性、大样本量数据，建立多层次ML模型探究AD的发病机制将是未来研究的重点.

关键词 阿尔茨海默病，遗传数据，机器学习
中图分类号 R592，TP39 **DOI**：10.16476/j.pibb.2020.0371

阿尔茨海默病（Alzheimer disease，AD）是一种进行性的、不可逆转的神经退行性疾病，伴随着记忆减退和日常沟通、活动障碍等症状[1]. 作为最常见的痴呆症类型，AD估计占所有痴呆症病例的60%~80%，通常始发于中老年，可能由神经元内和周围的蛋白质积累引起[2]. 作为一种常见的神经退行性疾病，全球AD患者已达2 000万以上，预计到2050年，全世界痴呆症患者的数量将达到1.315亿[3]，因此，研究治疗阿尔茨海默病已经迫在眉睫.

AD的常见分类，可根据发病年龄分为早发性AD（early onset AD，EOAD）和晚发性AD（late onset AD，LOAD），也可根据家族聚集性的存在，分为家族性AD和散发性AD. 通常AD病程进展中，轻度认知障碍（mild cognitive impairment，MCI）是从正常到AD过渡的一个重要环节，患者伴有认知能力的轻度变化，但仍能进行日常活动[4]. 其中，以记忆力丧失为主要症状的MCI转归为AD的风险显著增高，以每年约10%~15%的速度递进发展成为AD[5]. 大多数AD病例的病因尚不清楚，被认为是神经生物学和免疫学过程中涉及的遗传和环境因素之间复杂相互作用的结果，但估计70%的风险可归因于遗传因素[6]. 目前，淀粉样蛋白假说是AD发病机制的主流学说. 这一假说认为，各种因素促使淀粉样蛋白聚集形成不溶性斑块，导致β淀粉样蛋白在大脑中积聚，从而形成神经炎症和神经元内神经原纤维缠结，最终导致神经元功能障碍和死亡，但其病因尚不明确[7].

AD确诊的“金标准”是在病理上发现淀粉样斑块和神经原纤维缠结. 在临床上，AD的主要诊断方法一般是通过对患者进行神经心理学测验、脑脊液（cerebrospinal fluid，CSF）生物标记物、神经影像学检查或基因检查确诊[8]. 神经心理学测验

* 国家自然科学基金面上项目（61971275），国家自然科学基金重点项目（81830052），分子影像学重点实验室建设项目（18DZ2260400）资助.
** 通讯联系人.
Tel：13636316541，E-mail：yao6636329@hotmail.com
收稿日期：2020-12-12，接受日期：2021-01-11

易受主观因素和外部环境干扰，其检测结果可供临床参考；CSF生物标记物提取时需腰椎穿刺，侵入性较强，难以用于临床普查筛选和跟踪检测；神经影像学检查以正电子发射断层扫描（positron emission tomography，PET）和磁共振成像（magnetic resonance imaging，MRI）为主. 因PET能显示出脑内淀粉样蛋白斑块和神经纤维缠结，已被广泛应用于AD诊断，但存在特异性不强、分辨率较低、检测费用昂贵、操作复杂等缺点，难以普及. 结构MRI（structural MRI，sMRI）用于识别AD脑结构变化，对早期AD不敏感. 功能MRI（functional MRI，fMRI）主要检测脑功能区血流量的改变，但对患者体动敏感，易产生伪影[9]；基因检查主要以检测载脂蛋白E（apolipoprotein E，ApoE）基因为主，其他风险基因尚不明确，且价格较贵，难以普适[10]. 近年来也有采用单分子免疫阵列分析技术测量血液中神经丝轻链（neurofilament light chain，NFL）水平用于提前预知疾病的发生，为NFL作为AD临床诊断的生物标志物提供了可能性[11].

患者一旦确诊为AD，并无有效明确的治疗方案，但在早期MCI阶段进行物理治疗和药物干预，可有效延缓其病程发展. 物理治疗主要采用3种方法：光线疗法、经颅直流电刺激（transcranial direct current stimulation，tDCS）和重复经颅磁刺激疗法（transcranial magnetic stimulation，TMS）. 光线疗法对AD患者的睡眠障碍疗效较好，tDCS和TMS可改善AD患者的认知功能障碍，均效果有限[12]. 在药物干预上，开发AD新药物已成全球难题，美国食品药品监督管理局仅批准了4种药物（多奈哌、加兰他敏、利凡斯的明、美金刚）来改善临床症状[13]. 2019年我国首项原创治疗AD新药九期一[14]，根据国家药品上市许可持有人制度改革试点政策有条件批准上市，用于轻度至中度AD治疗，改善患者认知功能，填补了国内药物空白，有效地提高了我国AD患者的用药可适性.

目前，还没有预防和可逆治疗AD的解决方案，主要是因为AD涉及许多因素之间的复杂相互作用，而人类的复杂性使得无法使用简化的模型来理解AD. 机器学习（machine learning，ML）是人工智能的一种应用，它使系统能够自动学习并从经验中进行改进，而无需进行明确的编程. 近年来，ML技术的快速发展为解决涉及海量数据和超复杂结构性问题提供了契机. 在AD的研究中，研究数据主要分为遗传数据和非遗传数据. 非遗传数据包括神经影像学数据、生化生物学数据和神经生理学数据，目前对非遗传数据的研究已较为深入和普遍，如在AD的诊断和预后方面，ML技术利用神经影像学数据进行建模已占全部数据类型的近83.3%[15]. 较非遗传数据而言，对AD遗传数据的研究仍处于初步阶段. 而遗传因素作为大多数AD患者的主要病因，一直是AD发病机制研究的重点. 随着高通量测序技术[16]的广泛应用，利用遗传数据的研究成为热点. 这种情况迫切需要ML技术的助力. 因此，本文对该领域的研究进行了全面的回顾，主要介绍基于遗传数据的ML在AD中的应用，包括AD的诊断和预后、遗传变异分析和基因表达谱分析，并进行总结和展望.

1　ML在AD诊断与预后研究中的应用

目前尚没有治疗AD的方法，但在早期诊断出AD并加以干预能够有效延缓其发展进程，故准确诊断、预测AD对患者的治疗具有重要作用. ML通过对数据预处理、手动/自动特征提取和选择、目标模型构建和模型泛化能力评估这一流程，从给定的任务中学习潜在的数据模式来对AD进行诊断和预后. 由于数据模态的不同，对ML模型的泛化性有着较大的差异. 故本文将数据分为单模态和多模态来介绍基于遗传数据的ML在AD诊断和预后研究中的应用（表1）.

在AD研究中，通常利用患者的遗传信息来预测AD的诊断和预后. 在利用单模态的遗传数据包括基因型数据、基因表达谱等研究中，Oriol等[17]采用单核苷酸多态性（single nucleotide polymorphism，SNP）数据，比较最小绝对值收敛和选择算子（least absolute shrinkage and selection operator，LASSO）、K最近邻（k-nearest neighbor，KNN）、支持向量机（support vector machine，SVM）等ML算法预测LOAD，其中SVM为最高性能达到了接收者操作特征曲线下面积（area under curve，AUC）的72%. 除了用传统的遗传数据（基因型数据）来预测AD外，Xu等[18]创造性地采用两个连续氨基酸的频率用来描述序列信息，并基于SVM算法来分析基因编码的蛋白质序列，该方法识别AD的准确率为85.7%. 然而，这项研究的不足之处在于它没有区分早发性家族性AD和其

他类型AD之间的蛋白质序列信息. 而Castillo-Barnes等[19]考虑把影响常染色体显性遗传性AD的3个基因（PSEN1、PSEN2和APP）联合或单独作为实验数据，并以SVM为分类器，每个亚组与正常组之间的比较得出了大约80%的分类率，强调了将显性遗传性AD视为一个异构实体的重要性. 在分析基因表达数据上，Voyle等[20]采用递归特征消除随机森林（recursive feature elimination-random forest，RFE-RF），对血液基因表达数据建模，区分健康对照（health control，HC）、MCI和AD准确率为62.7%. Lee等[21]基于血液基因表达数据，利用深度神经网络（deep neural networks，DNN）来区分HC和AD，AUC值为0.859. Moradi等[22]采用线性判别分析（linear discriminant analysis，LDA）对血液基因表达谱进行分析，HC与AD分类的AUC值为0.84，HC与MCI分类的AUC值为0.80. 以上基因表达数据的研究表明，同时识别HC、MCI、AD比识别二元类别要困难许多. 有研究者用SNP数据，运用遗传算法（genetic algorithm，GA）、LASSO算法进行分步预测，AUC值提高了约5%，达到0.84[23]. 而在AD分类和预测上，最为普遍采用的是利用ML技术对神经影像数据进行建模，其对AD分类的平均准确率高达85%，对MCI向AD转化的预测准确率达83.7%[24]. 因此，这在仅使用遗传数据的情况下具有较强的竞争性.

在利用多模态数据上，与遗传数据结合的模态很多，以影像数据、临床数据、组学数据居多. Liu等[25]将弥散张量成像（diffusion tensor imaging，DTI）和SNP数据输入到深度卷积神经网络（deep convolutional neural networks，DCNN）中来预测AD，AUC值最高到0.858 3. Varol等[26]提出异质性判别分析（heterogeneity through discriminative analysis，HYDRA）算法，采用sMRI和SNP来预测AD，AUC值最高达0.942 3. 同样的，将sMRI和SNP数据结合，Ning等[27]采用神经网络（neural network，NN）预测AD，AUC值为0.992，而使用单一遗传数据的AUC值仅为0.689. Bi等[28]基于功能磁共振成像（functional magnetic resonance imaging，fMRI）和SNP，采用多模态随机森林算法预测AD准确率达到了89%. 这些研究表明，相比于单一遗传数据，结合影像数据对AD预测准确率有较好的提升. 在结合影像数据的基础上，添加临床数据作为补充的AD研究也有较好的表现. Gray等[29]基于sMRI、氟脱氧葡萄糖-正电子体层扫描（fluorodeoxyglucose positron emission tomography，FDG-PET）、CSF生物标记物和ApoE基因型，使用多模态随机森林预测AD准确率为83.3%. 同样的，Kohannim等[30]结合sMRI、FDG-PET、CSF生物标志物、ApoE基因型、年龄、性别和体重指数，选用SVM模型，用于MCI分类和预测未来（1年）认知功能下降，最高达90%的准确率. 为了区分稳定性和转换性MCI，Dukart等[31]用朴素贝叶斯（naive Bayesian，NB）算法，基于ApoE基因型、神经心理评估、sMRI、FDG-PET，准确率约87%，并且表明结合成像、遗传和/或神经心理评估可以比单一的模态分类器更可靠地区分稳定和转换的MCI. Spasov等[32]采用卷积神经网络（convolutional neural networks，CNN），对sMRI、人口统计学、神经心理评估和ApoE基因型数据建模，对3年内发展为AD的MCI和同一时间段内MCI稳定的患者进行分类，AUC值达0.925. Zhou等[33]用三阶段深度特征学习和融合同时预测HC、MCI和AD，准确率为0.65，但仍高于其他ML分类方法. 遗传信息除了与影像和临床信息进行联合使用，其与组学信息的联合也是AD研究的新兴热点方向. Shigemizu等[34]整合基因组数据和微RNA（microRNA）表达谱，构建了基于比例风险回归模型（proportional hazards model，又称COX模型）的预后预测模型来检测MCI向AD转换的高危个体，并在独立测试集上获得了0.702的一致性指数. Park等[35]为了结合基因表达和DNA甲基化数据，提出了基于差异表达基因和差异甲基化位置的特征选择方法，从而避免传统特征选择不能反映生物过程和无法保证减少的特征将保留其生物学意义的问题. 该方法首先确定差异表达基因和差异甲基化位置，再采用交叉法将两者相交进行整合，最后输入到贝叶斯优化的深度神经网络中，区分HC和AD准确率达到了0.823. 以上研究表明，基于遗传数据的ML方法对预测AD的预后和风险分层具有一定的价值，并且采用多模态数据如结合影像数据、临床数据以及组学数据，其准确性将大大提高[36-37].

Table 1　Application of machine learning based on genetic data in diagnosis and prognosis of AD

表1　基于遗传数据的机器学习在AD诊断和预后中的研究汇总

模态	数据类型	模型[1)]	样本数量[2)]	交叉验证法[3)]	研究结果	文献
单模态	SNP	LASSO、KNN、SVM	HC：371，AD：267	CV	最高分类性能SVM的AUC值为0.72	［17］
	基因编码的蛋白质序列	SVM	HC：1463，AD：279	—	识别AD的准确率为85.7%	［18］
	基因型数据	SVM	HC：173，AD：171	10-fold CV	识别AD准确率为80%	［19］
	基因表达	RFE-RF	HC：225，AD：218，MCI：305	独立验证集	同时预测HC、MCI和AD准确率为62.7%	［20］
	基因表达	DNN	HC：374，AD：347	10-fold CV	识别AD，AUC值为0.859	［21］
	基因表达谱	LDA	HC：239，AD：284，MCI：190	独立验证集	分类AD、HC的AUC值为0.84，分类MCI、HC的AUC值为0.80	［22］
	SNP	GA	HC：1017，AD：813	10-fold CV	识别AD，AUC值为0.84	［23］
多模态	SNP、DTI	DCNN	HC：100，AD：51	5-fold CV	预测AD，AUC值为0.858 3	［25］
	SNP、sMRI	HYDRA	HC：139，AD：103	—	预测AD，AUC值为0.942 3	［26］
	SNP、sMRI	NN	HC：225，AD：138，MCI：358	5-fold CV	预测AD，AUC值为0.992	［27］
	SNP、fMRI	MRF	HC：35，AD：37	—	预测AD准确率89%	［28］
	ApoE基因型、sMRI、CSF、FDG-PET	MRF	HC：35，AD：37，MCI：75	4-fold CV	预测AD准确率83.3%	［29］
	ApoE基因型、sMRI、CSF、FDG-PET、年龄、性别和体重指数	SVM	HC：213，AD：158，MCI：264	LOOCV	分类MCI和预测未来（1年）认知功能下降，达到90%的准确率	［30］
	ApoE基因型、sMRI、FDG-PET、神经心理评估	NB	HC：112，AD：144，sMCI：265，cMCI：177	独立验证集	准确率87%，表明结合成像、遗传信息比单模态更能区分稳定和转换MCI	［31］
	ApoE基因型、sMRI、神经心理评估、人口统计	CNN	HC：184，AD：192，sMCI：228，cMCI：181	10-fold CV	对3年内发展为AD的MCI和MCI稳定的患者分类，AUC值为0.925	［32］
	SNP、sMRI、FDG-PET	DFFF	HC：226，AD：190，MCI：389	20-fold CV	预测HC、MCI、AD准确率为0.65	［33］
	microRNA表达谱、SNP	COX	HC：91，AD：271，MCI：248	CV	检测MCI向AD转换的高危个体，获得0.702一致性指数	［34］
	基因表达和DNA甲基化	DNN	HC：257，AD：439	5-fold CV	准确率0.823，证明结合基因表达和DNA甲基化数据可提高预测准确性	［35］
	SNP、sMRI、FDG-PET、CSF	SMML	HC：47，AD：49，MCI：93	10-fold CV	AD vs MCI vs HC最佳准确率为0.71	［36］
	SNP、sMRI、FDG-PET、神经心理评估	GBM	HC：301，AD：252，MCI：471	10-fold CV	预测AD，AUC值为0.876	［37］

[1)] MRF：多模态随机森林（multimodal random forest）；DFFF：深度特征学习和融合（deep feature learning and fusion framework）；COX：比例风险回归模型；SMML：稀疏多模态学习（sparse multi model learning）；GBM：梯度提升机（gradient boosting machines）. [2)] HC：健康对照（health control）组；AD：阿尔茨海默病（Alzheimer's disease）患者组；MCI：轻度认知障碍（mild cognitive impairment）组；sMCI：稳定型轻度认知障碍（stable mild cognitive impairment）组；cMCI：MCI转化为AD（MCI converting to AD）组. [3)] CV：交叉验证法（cross validation）；10-fold CV：10折交叉验证（10-fold cross-validation）；5-fold CV：5折交叉验证（5-fold cross-validation）；4-fold CV：4折交叉验证（4-fold cross-validation）；LOOCV：留一交叉验证（leave-one-out-cross-validation）；20-fold CV：20折交叉验证（20-fold cross-validation）.

2　ML在AD遗传变异分析中的应用

AD的遗传变异数据具有高维性，并且这些遗传变异的大多数与AD无关. 目前，遗传关联研究已经揭示了一些潜在的AD易感基因，但仍有必要确定未知的AD相关基因和治疗靶点，以便更好地了解AD的致病机制，从而开发有效的AD治疗方法.

ML在AD遗传变异分析中的研究仍处于探索阶段，但已有一些较好的成果（表2）. 为了筛选AD相关的SNP，Abd等[38]利用SNP数据，分别采用序列最小最优化（sequential minimal optimization，SMO）算法、NB、树增强朴素贝叶斯（tree augmented naive Bayes，TANB）和K2学习算法，检测出500个SNP. 在探究遗传变异与表型特征之间的关联上，Wang等[39]将网络体素信息和网络连通性信息作为连接遗传风险因素和疾病状态的中间特征，利用诊断对齐多模态回归（diagnosis-aligned multi-modality regression，DAMM）方法，证明了连通性特征可以作为识别表型标记的补充信息. Vounou等[40]对SNP和拷贝数变异（copy number variation，CNV）数据，使用稀疏降秩回归（sparse reduced-rank regression，SRRR），证实了ApoE和线粒体外膜转位酶40（translocase of outer mitochondrial membrane 40，TOMM40）基因的关键作用. Hao等[41]使用树引导的稀疏学习（tree-guided sparse learning，T-GSL）来识别SNP和sMRI的关联，结果筛选出10个SNP与左、右侧海马区结构体积变化相关性最强，与晚发性AD风险显著相关. 从上述论文的结果表明，使用ML算法分析SNP数据可以发现可能与AD风险相关的新基因[42-43].

Table 2　Application of machine learning based on genetic data in genetic variation analysis of AD

表2　基于遗传数据的机器学习在AD遗传变异分析中的研究汇总

数据类型	模型[1)]	样本数量[2)]	交叉验证法[3)]	研究结果[4)]	文献
SNP	SMO、TANB	HC：214，AD：177，MCI：366	10-fold CV	检测出500个SNP	［38］
ApoE基因型	DAMM	HC：38，AD：26，SMC：19，EMCI：40，LMCI：34	5-fold CV	壳核灰质的减少与APOE SNP rs429358相关	［39］
SNP、CNV	SRRR	HC：153，AD：101，sMCI：114，cMCI：107	CV	证实了ApoE和TOMM40的关键作用	［40］
SNP	T-GSL	HC：210，AD：173，MCI：360	NCV	筛选出10个SNP与左、右侧海马区结构体积变化相关性最强，与LOAD显著相关	［41］
SNP	SRRR	MCI：189	独立验证集	相比MULM，SRRR能更好检测有害遗传变异	［42］
SNP	T-SCCA	HC：211，AD：160，MC：82，EMCI：273，LMCI：187	5-fold CV	T-SCCA在发现多重关联方面优于SCCA	［43］

[1)] TANB：树增强朴素贝叶斯（tree augmented naive bays）. [2)] HC：健康对照（health control）组；AD：阿尔茨海默病（Alzheimer's disease）患者组；MCI：轻度认知障碍（mild cognitive impairment）组；SMC：主观记忆担忧（significant memory concern）组；EMCI：早期轻度认知障碍（early mild cognitive impairment）组；LMCI：晚期轻度认知障碍（late mild cognitive impairment）组；sMCI：稳定型轻度认知障碍（stable mild cognitive impairment）组；cMCI：MCI转化为AD（MCI converting to AD）组. [3)] 10-fold CV：10折交叉验证（10-fold cross-validation）；5-fold CV：5折交叉验证（5-fold cross-validation）；CV：交叉验证法（cross validation）；NCV：嵌套交叉验证（nested cross validation）. [4)] MULM：质量-单变量线性建模（mass-univariate linear modelling）；T-SCCA：三向稀疏典型相关分析（three-way sparse canonical correlation analysis）；SCCA：稀疏典型相关分析（sparse canonical correlation analysis）.

3　ML在AD基因表达谱分析中的应用

遗传变异单独或与环境因素联合可改变脑细胞的基因表达谱，导致某些蛋白质代谢异常，最终导致AD的病理改变. 研究脑细胞基因表达水平的变化有助于发现与AD发病相关的关键基因和途径，可能成为治疗干预的靶点.

由于基因表达谱数据的高维性，许多研究已经从传统的统计方法转向ML进行数据分析，并有效揭示了AD的生物学特性（表3）. Wang等[44]整合

6个脑区的基因表达谱，构建差异共表达网络，并使用SVM分类，得到了44个以网络模块形式被定位为潜在生物标志物的基因. Kong等[45]基于DNA微阵列基因表达数据和相应的基因聚类，使用独立成分分析（independent component analysis，ICA），提取了高表达水平的50多个重要基因. 在脑细胞基因表达水平变化的同时，其蛋白质和RNA水平也随之改变. 针对这一生物特性，Gutiérrez等[46]通过决策树（decision tree，DT）来分析69个基因的表达以及APP、ApoE、BACE1、NCSTN的表达水平，实现了对tau蛋白表达水平的鉴定. Zhou等[47]用径向基核支持向量机（radial basis function-svm，RBF-SVM）算法，对脑老化和AD四个脑区的长非编码RNA（lncRNA）表达谱进行了比较分析. 分析表明，lncRNA表达变异参与了脑发育和代谢相关的生物学过程，强调了lncRNA在AD发病机制中的重要性. 在探索AD发病相关通路方面，Zhang等[48]提出一种多尺度网络建模方法（multiscale network modeling approach，MNMA），将基因表达和基因型分析用贝叶斯网络进行因果推理，确定了与晚发性AD有因果联系的子网. Martínez-Ballesteros等[49]通过DT，定量关联规则（quantitative association rule，GAR）和层次聚类（hierarchical clustering，HC）来集成分析基因表达谱，发现90个基因在AD患者中发生了显著改变. Kong等[50]将ICA和非负矩阵分解（nonnegative matrix factorization，NMF）相结合对基因表达谱进行分析，识别了1 500多个关键基因. 利用基因型数据来识别AD基因，不能区分多效性和因果中介作用，并会受到数据的强假设限制. 为了克服这些局限性，Park等[51]开发了一种贝叶斯推理框架（causal multivariate mediation within extended linkage disequilibrium，CaMMEL），基于SNP和基因表达数据，发现了206个致病基因. 以上研究表明，使用ML来分析基因表达谱可以帮助发现在AD发病机制中发挥重要作用的基因和途径[52-55].

Table 3　Application of machine learning based on genetic data in gene expression profile analysis of AD

表3　基于遗传数据的机器学习在AD基因表达谱分析中的研究汇总

数据类型	模型[1)]	样本数量[2)]	交叉验证法[3)]	研究结果[4)]	文献
基因表达谱	SVM	HC：74，AD：87	LOOCV	得到44个潜在生物标志物基因	［44］
基因表达谱	ICA	HC：8，AD：5	独立验证集	提取了高表达的50多个基因	［45］
基因表达谱	DT	HC：9，AD：22	10-fold CV	鉴定了tau蛋白的表达水平	［46］
基因表达谱	RBF-SVM	HC：131，AD：144	独立验证集	lncRNA表达变异影响脑发育和代谢	［47］
基因表达谱	MNMA	HC：273，AD：376	—	确定AD因果联系的子网	［48］
基因表达谱	DT、GAR、HC	HC：13，AD：20	5-fold CV	发现显著改变的90个基因	［49］
基因表达谱	ICA和NMF	HC：8，AD：5	独立验证集	识别了1 500多个关键基因	［50］
SNP、基因表达	CaMMEL	HC：25 580，AD：48 466	独立验证集	发现206个致病基因	［51］
氨基酸频率	DT、LR、SVM等11种	AD基因：458，HC基因：55 947	5-fold CV	识别13个新的候选基因	［52］
基因表达	SVM	AD基因：335，HC基因：22 646	独立验证集	候选基因分类准确率84.56%，AUC值为0.94	［53］
DNA甲基化表达	RF	HC：34，AD：151	5-fold CV	确定MYNN为最佳生物标志物	［54］
基因表达	TSCC	HC：9，AD：22	5-fold CV	识别了13个潜在AD相关基因	［55］

[1)] LR：逻辑回归（logistic regression）；RF：随机森林（random forest）；TSCC：两级级联分类器（two-stage cascaded classifier）. [2)] HC：健康对照（health control）组；AD：阿尔茨海默病（Alzheimer's disease）患者组. [3)] LOOCV：留一交叉验证（leave-one-out-cross-validation）；10-fold CV：10折交叉验证（10-fold cross-validation）；5-fold CV：5折交叉验证（5-fold cross-validation）. [4)] MYNN：肌神经蛋白（myoneurin）.

4　总结与展望

开发技术的计算能力和数据挖掘能力每年呈指数级增长. 这些新技术使针对复杂生物过程和疾病的分析成为可能，这些过程和疾病具有非同寻常的规模和多个维度. 特别是对于复杂的疾病，包括

AD，单一或几个维度的分析使人们无法捕捉到与这些疾病相关的确切原因和因素. 因此，必须使用高效但复杂的方法来组合多种数据类型以精确定位疾病的特定因素. 目前，用于ML研究的数据库相对有限，很少有研究在基因、蛋白质、新陈代谢和环境因素的多个水平上进行综合分析. 大多数研究的结论只是为进一步研究提供参考，很少有研究对这些发现进行生物学验证或提出验证方案. 同时，也少有研究根据所获得的结果提出AD发病机制的理论框架.

大多数AD病例是遗传和环境因素复杂交互作用的结果，传统的遗传分析方法发现了许多AD发病的重要遗传因素. 近年来，随着高通量测序技术发展，传统的统计方法对遗传数据的分析已显示出一定的局限性. 近10年来，基于遗传数据的ML已被应用于AD诊断和预后研究，以及遗传变异和基因表达谱分析，并取得了一些有意义的结果：利用ML方法对预测AD的预后和风险分层具有一定价值，并且采用多模态数据如结合影像数据、临床数据、组学数据，其准确性将大大提高；使用ML分析SNPS数据可以发现可能与AD风险相关的新基因；ML分析基因表达谱可以帮助发现在AD发病机制中发挥重要作用的基因和途径. 虽然这些还处于比较初步的阶段，但随着高质量、综合性、大样本量的数据不断增加，多层次ML模型的开发，以及计算能力的惊人进步，最终将发展出一套全面的分析体系，有助于全面了解AD的发病机制.

参 考 文 献

[1] Liu X, Hou D R, Lin F B, *et al*. The role of neurovascular unit damage in the occurrence and development of Alzheimer's disease. Reviews in the Neurosciences, 2019, **30**(5): 477-484

[2] Falahati F, Westman E, Simmons A. Multivariate data analysis and machine learning in Alzheimer's disease with a focus on structural magnetic resonance imaging. Journal of Alzheimer's disease, 2014, **41**(3): 685-708

[3] Sallim A B, Sayampanathan A A, Cuttilan A, *et al*. Prevalence of mental health disorders among caregivers of patients with Alzheimer disease. Journal of The American Medical Directors Association, 2015, **16**(12): 1034-1041

[4] Piaceri I, Nacmias B, Sorbi S. Genetics of familial and sporadic Alzheimer's disease. Frontiers in Bioscience (Elite edition), 2013, **5**(1): 167-177

[5] Grundman M, Petersen R C, Ferris S H, *et al*. Mild cognitive impairment can be distinguished from Alzheimer disease and normal aging for clinical trials. Archives of Neurology, 2004, **61**(1): 59-66

[6] Freudenberg-Hua Y, Li W, Davies P. The role of genetics in advancing precision medicine for Alzheimer's disease—a narrative review. Frontiers in Medicine, 2018, **5**: 108

[7] Lane C A, Hardy J, Schott J M. Alzheimer's disease. European Journal of Neurology, 2018, **25**(1): 59-70

[8] Giacobini E, Gold G. Alzheimer disease therapy—moving from amyloid- β to tau. Nature Reviews Neurology, 2013, **9**(12): 677-686

[9] 黄 蕊, 华学思, 张 兰. 影像学技术应用于阿尔茨海默病早期诊断的研究进展. 中国康复理论与实践, 2017, **23**(5): 534-538
Huang R, Hua X S, Zhang L. Chinese Journal of Rehabilitation Theory and Practice, 2017, **23**(5): 534-538

[10] 徐 庆, 姚 芳, 沈立明, 等. 阿尔茨海默病血液生物标志物新进展及预测. 阿尔茨海默病及相关病, 2019, **2**(3): 444-449
Xu Q, Yao F, Shen L M, *et al*. Chinese Journal of Alzheimer's Disease and Related Disorders, 2019, **2**(3): 444-449

[11] Oliver P, Schultz S A, Apel A, *et al*. Serum neurofilament dynamics predicts neurodegeneration and clinical progression in presymptomatic Alzheimer's disease. Nature Medicine, 2019, **25**(2): 277-283

[12] 杨金菊, 邹显巍, 余建萍, 等. 阿尔茨海默病物理治疗的作用机制及应用. 国际老年医学杂志, 2020, **41**(4): 262-265
Yang J G, Zou X W, Yu J P, *et al*. International Journal of Geriatrics, 2020, **41**(4): 262-265

[13] Martins M, Silva R, M M Pinto M, *et al*. Marine natural products, multitarget therapy and repurposed agents in Alzheimer's disease. Pharmaceuticals, 2020, **13**(9): 242

[14] 用于治疗轻、中度阿尔茨海默症国家1类新药甘露特纳胶囊获批上市. 开卷有益-求医问药, 2019, **26**(12): 4-5
Ganlutna capsule, a national class 1 new drug for the treatment of mild to moderate Alzheimer's disease, has been approved to be put on the market. Journal for Beneficial Readines Drug Information & Medical Advices, 2019, **26**(12): 4-5

[15] Tanveer M, Richhariya B, Khan R U, *et al*. Machine learning techniques for the diagnosis of Alzheimer's disease: A review. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2020, **16**(1s): 1-35

[16] Fenoglio C, Scarpini E, Serpente M, *et al*. Role of genetics and epigenetics in the pathogenesis of Alzheimer's disease and frontotemporal dementia. Journal of Alzheimer's Disease, 2018, **62**(3): 913-932

[17] Oriol J D V, Vallejo E E, Estrada K, *et al*. Benchmarking machine learning models for late-onset Alzheimer's disease prediction from genomic data. BMC Bioinformatics, 2019, **20**(1): 1-17

[18] Xu L, Liang G, Liao C, *et al*. An efficient classifier for Alzheimer's disease genes identification. Molecules, 2018, **23**(12): 3140

[19] Castillo-Barnes D, Su L, Ramírez J, *et al*. Autosomal dominantly inherited Alzheimer disease: analysis of genetic subgroups by machine learning. Information Fusion, 2020, **58**: 153-167

[20] Voyle N, Keohane A, Newhouse S, *et al*. A pathway based classification method for analyzing gene expression for Alzheimer's disease diagnosis. Journal of Alzheimer's Disease,

2016, **49**(3): 659-669

[21] Lee T, Lee H. Prediction of Alzheimer's disease using blood gene expression data. Scientific Reports, 2020, **10**(1): 3485

[22] Moradi E, Marttinen M, Häkkinen T, *et al*. Supervised pathway analysis of blood gene expression profiles in Alzheimer's disease. Neurobiology of Aging, 2019, **84**: 98-108

[23] Romero-Rosales B L, Tamez-Pena J G, Nicolini H, *et al*. Improving predictive models for Alzheimer's disease using GWAS data by incorporating misclassified samples modeling. PLoS One, 2020, **15**(4): e0232103

[24] Jo T, Nho K, Saykin A J. Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data. Frontiers in Aging Neuroscience, 2019, **11**: 220

[25] Liu Y, Li Z, Ge Q, *et al*. Deep feature selection and causal analysis of Alzheimer's disease. Frontiers in Neuroscience, 2019, **13**: 1198

[26] Varol E, Sotiras A, Davatzikos C, *et al*. HYDRA: revealing heterogeneity of imaging and genetic patterns through a multiple max-margin discriminative analysis framework. NeuroImage, 2017, **145**(Pt B): 346-364

[27] Ning K, Chen B, Sun F, *et al*. Classifying Alzheimer's disease with brain imaging and genetic data using a neural network framework. Neurobiology of Aging, 2018, **68**: 151-158

[28] Bi X, Liu Y, Wang Y, *et al*. Effective diagnosis of Alzheimer's disease via multimodal fusion analysis framework. Frontiers in Genetics, 2019, **10**: 976

[29] Gray K R, Aljabar P, Heckemann R A, *et al*. Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. NeuroImage, 2013, **65**: 167-175

[30] Kohannim O, Hua X, Hibar D P, *et al*. Boosting power for clinical trials using classifiers based on multiple biomarkers. Neurobiology of Aging, 2010, **31**(8): 1429-1442

[31] Dukart J, Sambataro F, Bertolino A. Accurate prediction of conversion to Alzheimer's disease using imaging, genetic, and neuropsychological biomarkers. Journal of Alzheimer's Disease, 2016, **49**(4): 1143-1159

[32] Spasov S, Passamonti L, Duggento A, *et al*. A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. NeuroImage, 2019, **189**: 276-287

[33] Zhou T, Thung K H, Zhu X, *et al*. Feature learning and fusion of multimodality neuroimaging and genetic data for multi-status dementia diagnosis//Wang Q, Shi Y, Suk H I, Suzuki K. Machine Learning in Medical Imaging. Cham: Springer, 2017: 132-140

[34] Shigemizu D, Akiyama S, Higaki S, *et al*. Prognosis prediction model for conversion from mild cognitive impairment to Alzheimer's disease created by integrative analysis of multi-omics data. Alzheimer's Research & Therapy, 2020, **12**(1): 145

[35] Park C, Ha J, Park S. Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset. Expert Systems with Applications, 2020, **140**: 112873

[36] Zhang Z, Huang H, Shen D, *et al*. Integrative analysis of multi-dimensional imaging genomics data for Alzheimer's disease prediction. Frontiers in Aging Neuroscience, 2014, **6**: 260

[37] Khanna S, Domingo-Fernández D, Iyappan A, *et al*. Using multi-scale genetic, neuroimaging and clinical data for predicting Alzheimer's disease and reconstruction of relevant biological mechanisms. Scientific Reports, 2018, **8**(1): 11173

[38] Abd El Hamid M M, Mabrouk M S, Omar Y M K. Developing an early predictive system for identifying genetic biomarkers associated to Alzheimer's disease using machine learning techniques. Biomedical Engineering: Applications, Basis and Communications, 2019, **31**(5): 1950040

[39] Wang M, Hao X, Huang J, *et al*. Discovering network phenotype between genetic risk factors and disease status *via* diagnosis-aligned multi-modality regression method in Alzheimer's disease. Bioinformatics, 2019, **35**(11): 1948-1957

[40] Vounou M, Janousova E, Wolz R, *et al*. Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. NeuroImage, 2012, **60**(1): 700-716

[41] Hao X, Yu J, Zhang D. Identifying genetic associations with MRI-derived measures via tree-guided sparse learning//Golland P, Hata N, Barillot C, Hornegger J, Howe R. Medical Image Computing and Computer-assisted Intervention. Cham: Springer, 2014: 757-764

[42] Vounou M, Nichols T E, Montana G, *et al*. Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank regression approach. NeuroImage, 2010, **53**(3): 1147-1159

[43] Hao X, Li C, Du L, *et al*. Mining outcome-relevant brain imaging genetic associations *via* three-way sparse canonical correlation analysis in Alzheimer's disease. Scientific Reports, 2017, **7**: 44272

[44] Wang L, Liu Z P. Detecting diagnostic biomarkers of Alzheimer's disease by integrating gene expression data in six brain regions. Frontiers in Genetics, 2019, **10**: 157

[45] Kong W, Mou X, Liu Q, *et al*. Independent component analysis of Alzheimer's DNA microarray gene expression data. Molecular Neurodegeneration, 2009, **4**(1): 5

[46] Gutiérrez S L M, Rivero M H, Ramírez N C, *et al*. Decision trees for the analysis of genes involved in Alzheimer’s disease pathology. Journal of Theoretical Biology, 2014, **357**: 21-25

[47] Zhou M, Zhao H, Wang X, *et al*. Analysis of long noncoding RNAs highlights region-specific altered expression patterns and diagnostic roles in Alzheimer's disease. Briefings in Bioinformatics, 2019, **20**(2): 598-608

[48] Zhang B, Tran L, Emilsson V, *et al*. Characterization of genetic networks associated with Alzheimer's disease//Castrillo J, Oliver S. Systems Biology of Alzheimer's Disease. New York: Humana Press, 2016: 459-477

[49] Martínez-Ballesteros M, García-Heredia J M, Nepomuceno-Chamorro I A, *et al*. Machine learning techniques to discover genes with potential prognosis role in Alzheimer's disease using different

biological sources. Information Fusion, 2017, **36**: 114-129

[50] Kong W, Mou X, Hu X. Exploring matrix factorization techniques for significant genes identification of Alzheimer's disease microarray gene expression data. BMC Bioinformatics, 2011, **12**(5): 7

[51] Park Y, Sarkar A K, He L, *et al*. A Bayesian approach to mediation analysis predicts 206 causal target genes in Alzheimer's disease. bioRxiv, 2017, **12**(1): 219428

[52] Jamal S, Goyal S, Shanker A, *et al*. Integrating network, sequence and functional features using machine learning approaches towards identification of novel Alzheimer genes. BMC Genomics, 2016, **17**(1): 807

[53] Huang X, Liu H, Li X, *et al*. Revealing Alzheimer's disease genes spectrum in the whole-genome by machine learning. BMC Neurology, 2018, **18**(1): 5

[54] Ren J, Zhang B, Wei D, *et al*. Identification of methylated gene biomarkers in patients with Alzheimer's disease based on machine learning. BioMed Research International, 2020, **2020**: 8348147

[55] Miao Y, Jiang H, Liu H, *et al*. An Alzheimers disease related genes identification method based on multiple classifier integration. Computer Methods and Programs in Biomedicine, 2017, **150**: 107-115

Application of Machine Learning Based on Genetic Data in The Study of Alzheimer's Disease*

JIN Yu[1,2], YAO Xu-Feng[2,3]**, HAN Li-Ting[1,2], ZHAO Cong-Yi[1,2], HUANG Gang[2,3]

([1])*School of Medical Instrument and Food Engineering, University of Shanghai for Science and Technology, Shanghai* 200082, China;

[2])*College of Medical Imaging, Shanghai University of Medicine and Health Sciences, Shanghai* 201308, China;

[3])*Shanghai Key Laboratory of Molecular Imaging, Shanghai University of Medicine and Health Sciences, Shanghai* 201308, China)

Abstract　Alzheimer's disease (AD), a neurodegenerative disease, is closely related to the genetic and environmental factors, about 70% of which are caused by the genetic factors, but its pathogenesis is still unclear. Along with the advent of high-throughput gene sequencing, the processing of genetic data using machine learning (ML) has become a hot spot. In this paper, the applications of ML in AD are mainly reviewed, including the diagnosis and prognosis of AD based on genetic data, the analysis of genetic variation of AD, the analysis of gene expression profile of AD, and the further development of ML for AD. Firstly, during the diagnosis and prognosis of AD, the genetic data combining with other modalities, such as imaging data, clinical data and histological data, would be greatly improved the accuracy of ML methods. It is valuable for the early diagnosis of AD, and effectively delays the progression of AD. Secondly, the application of ML in the analysis of genetic variation of AD, single nucleotide polymorphisms (SNPs) of new genes were dug out, and the pathogenic mechanism of AD was further explored. Thirdly, the analysis of gene expression profile of AD mainly focuses on the discovery of the pathways of genes which could provide the possibility of gene targets for AD therapy. In the future, the multi-level model of ML might be developed for high-quality, diverse and large data, and provide scientific strategies for exploring the pathogenesis of AD.

Key words　Alzheimer's disease, genetic data, machine learning

DOI: 10.16476/j.pibb.2020.0371

---

* This work was supported by grants from The National Natural Science Foundation of China（61971275，81830052），Key Laboratory Construction Program of Molecular Imaging（18DZ2260400）.

** Corresponding author.

Tel：86-13636316541，E-mail：yao6636329@hotmail.com

Received：December 12，2020　Accepted：January 11，2021