doi: 10. 3969 / j. issn. 1674 - 1242. 2023. 01. 003

基于多层感知机的 DNA 甲基化年龄预测模型

宗西增[1]，蔡蕊蕊[2]，田若婷[1]，赵舜琳[2]，张黎[1]
（1. 长春工业大学计算机科学与工程学院，吉林长春 130012;
2. 东北师范大学信息科学与技术学院，吉林长春 130017）

【摘要】 衰老的过程中伴随着 DNA 甲基化的变化，DNA 甲基化成为重要的衰老生物标志物之一。近年来，人们对衰老领域的研究越发火热，年龄预测有助于研究生物衰老问题，但预测精度还有待进一步提高。以往的研究大多基于线性回归模型，使用 DNA 甲基化数据中与年龄高度相关的 CpG 位点作为特征进行年龄预测。相比机器学习模型，使用深度学习模型对多特征任务包容性更强，能够选取更多的 CpG 位点作为特征。在 Illumina 27K 和 Illumina 450K 阵列的甲基化数据中，选择共同的 21 368 个 CpG 位点的甲基化数据作为输入，使用多层感知机建立泛组织年龄预测方法 MLPAge 对年龄进行预测，将 MLPAge 与泛组织年龄预测方法行业中的标准 Horvath 353 CpG 时钟进行了比较。在来自 8 项研究的 2 310 个样本的独立验证集中，其绝对中位数误差（MAD）为 3.77 年。研究发现，多层感知机能够更好地提取与年龄相关的特征，在年龄预测方面具有更高的准确度，为该领域提供了一种新的基于深度学习的方案。

【关键词】 DNA 甲基化；CpG 位点；多层感知机；年龄预测
【中图分类号】Q811.4 【文献标志码】A 文章编号：1674-1242(2023)01-0034-08

DNA Methylation Age Prediction Model Based on Multilayer Perceptron

ZONG Xizeng[1], CAI Ruirui[2], TIAN Ruoting[1], ZHAO Shunlin[2], ZHANG Li[1]
（1. School of Computer Science and Engineering, Changchun University of Technology, Changchun, Jilin 130012, China;
2. College of Information Science and Technology, Northeast Normal University, Changchun, Jilin 130017, China）

【Abstract】 The aging process is accompanied by changes in DNA methylation that has become one of the important biomarkers of aging. Research in the field of aging has become more and more hot in recent years. Age prediction helps study biological aging the research of biological aging, but the prediction accuracy needs to be further improved. Most of the previous studies were based on linear regression models that used highly correlated CpG loci in DNA methylation data as features to predict age. Compared with machine learning models, deep learning models is more inclusive for multi-feature tasks and can select more CpG loci as features. In the methylation data of Illumina 27K and Illumina 450K arrays, methylation data of 21 368 CpG sites were selected as input. MLPAge, a pan-tissue age prediction method, was established using multi-layer perceptron to predict age. MLPAge

收稿日期：2023-03-01
基金项目：吉林省科技发展计划学科布局项目（20210101175JC）；吉林省教育厅“十三五”科学技术研究规划项目（JJKH20191309KJ）；吉林省发改委产业技术研究与开发项目（2022C043-2）。
作者简介：宗西增（1997—），男，山东省枣庄市人，硕士研究生，从事生物信息学研究。
通讯作者：张黎，女，讲师，硕士生导师，电话（Tel.）：0186-43128995，E-mail：lizhang@ccut.edu.cn。

was compared with the Horvath 353 CpG clock, the standard in the pan-tissue age prediction methods, industry in an independent validation set of 2 310 samples from 8 studies with Median Absolute Deviation (MAD) of 3.77 years. It was found that the multi-layer perceptron is able to better extract age-related features and has higher accuracy in age prediction, providing a new deep learning-based solution for the field.

【**Key words**】 DNA Methylation; CpG Loci; Convolutional Neural Network; Age Prediction

0　引言

衰老的过程中总是伴随着 DNA 甲基化的变化，DNA 甲基化和衰老息息相关[1]。衰老是癌症和退行性疾病的重要影响因素。为了更好地研究衰老机制，近年来出现了基于 DNA 甲基化的年龄预测方法，称为“表观遗传时钟”。这些年龄预测方法基于少量的 DNA 序列中 CpG 位点的甲基化数据，即可较为准确地预测生物年龄[2]。但是，目前存在的年龄预测方法大多基于线性回归模型，而线性回归模型只适用于线性关系，DNA 甲基化与年龄并不是单纯的线性关系，因此可能会出现高误差。本实验收集了 27 个公开可用的 DNA 甲基化数据集，训练开发了 MLPAge，使用来自 Illumina 27K 和 Illumina 450K 阵列的共同的 21 368 个 CpG 位点的甲基化数据进行年龄预测。MLPAge 使用多层感知机进行建模，对 21 368 个 CpG 位点进行特征筛选并提取特征之间的相互作用，降低了预测的误差。在独立数据集测试中，与当前的表观遗传时钟相比，MLPAge 具有更高的准确度和可解释性。

1　介绍

1.1　DNA 甲基化与衰老

衰老是生物体正常且必然的生理现象，随着时间的推移，生物体的各种机能逐渐衰退。衰老是基因、环境等多种因素导致的复杂现象。随着生物体的衰老，各种疾病发生的概率直线上升，如心血管疾病、退行性疾病和癌症[2]。虽然衰老的自然规律无法违背，但是延缓衰老和降低衰老带来的各种疾病的发生概率是当前生物学研究的热门领域。

DNA 甲基化是一个生物过程，它会在 DNA 分子中引入甲基化基团（$-CH_3$），但是甲基化并不会改变序列本身，而会改变 DNA 片段的活性[3]。当甲基化位于基因启动子区域时，DNA 甲基化通常起到抑制基因转录的作用。DNA 甲基化常发生于 DNA 序列中胞嘧啶（C）-磷酸（p）-鸟嘌呤（G）结构上，这种结构被称为 CpG 位点。最重要的一种 DNA 甲基化修饰就是 5-甲基-胞嘧啶（5mC）[4]，是在 DNA 甲基转移酶（DNMT）的作用下将甲基化基团添加到胞嘧啶的 5′C 位置上形成的。大量研究表明，衰老的过程中伴随着 DNA 甲基化的变化，从 DNA 甲基化角度去探究衰老机制值得深入研究，以便找到延缓甚至逆转衰老的途径。因此，DNA 甲基化，特别是 5-甲基-胞嘧啶（5mC），已成为预测生物年龄最有效的生物标记物之一[5]。

随着微阵列技术的快速发展，越来越多的 DNA 甲基化位点数据被研究人员所了解。Illumina 公司推出的 Infinium Human Methylation 27 Beadchip 包含 27 578 个 CpG 位点的甲基化数值，简称 27K。同样，450K 包含 485 577 个 CpG 位点的甲基化数值。对于探针的甲基化水平，beta 值是最佳的衡量标准[6]，beta 值的计算公式如下：

$$\text{beta} = \text{Methylation} / \left(\text{Methylation} + \text{Unmethylation} + \text{offset}\right)$$

式中，Unmethylation 代表 CpG 位点的非甲基化信号强度，Methylation 代表 CpG 位点的甲基化信号强度，为了防止分母为 0，额外设置偏移量 offset，通常将 offset 设置为 100，本文数据集处理均设置偏移量为 100。每个 CpG 位点上都具有相应的 beta 值，显示了该位点的 DNA 甲基化的程度，beta 值的范围是从 0（完全未甲基化）到 1（完全甲基化）。其中，beta 值大于或等于 0.6 表示该位点完全甲基化，beta 值小于或等于 0.2 表示该位点完全未甲基化，beta 值介于 0.2 和 0.6 之间表示该位点部分甲基化。

1.2　表观遗传时钟进展

基于 DNA 甲基化的年龄预测方法，称为表观遗传时钟。2013 年发表了两篇关于表观遗传时钟的开创性文章：Horvath 的 353 CpG 时钟[7]和 Hannum 等的 71 CpG 时钟[8]。这两种表观遗传时钟都描述了一种算法，即使用 DNA 甲基化的数据来预测人类的实际年龄。这

两种表观遗传时钟都依赖弹性净正则化回归方法，这是一种线性模型。其中特定的 CpG 位点的甲基化水平被分配权重，然后相加以获得最终预测的年龄值。其中 Horvath 的 353 CpG 时钟的中位绝对误差（MAD）为 3.6 年，Hannum 等的 71 CpG 时钟的均方根误差（RMSE）为 3.9 年。这两种表观遗传时钟在 DNA 甲基化年龄预测上都取得了良好的效果，也证实了年龄变化可以在 DNA 甲基化中表现出来。

Horvath 的 353 CpG 时钟选取了 27K 和 450K 阵列甲基化数据中共有的 353 个与年龄变化相关的 CpG 位点，通过将 8 000 多个不同组织和不同疾病的样本中 353 个 CpG 位点的 DNA 甲基化数据放入弹性网络中进行加权，得出每个 CpG 的权重信息和回归方程，进行年龄预测。Horvath 的方法在 DNA 甲基化年龄预测领域被多次引用，它被称为“目前最先进的泛组织表观遗传时钟”。本文也将其作为对比实验。Hannum 等的 71 CpG 时钟采集了 656 名年龄在 19 ~ 101 岁的人类个体全血样本。该方法在 450K 数据的 485 577 个 CpG 位点的基础上去除了性染色体包含的 CpG 位点，对剩余的 CpG 位点数据进行回归，采用弹性网络的惩罚多变量回归方法建立年龄预测模型，最佳模型选择了一组 71 个高度预测年龄的甲基化 CpG 位点。

Horvath 和 Hannum 等的成功，使越来越多的研究人员关注 DNA 甲基化年龄预测领域，并启发其他研究人员使用相同的理念开发更多的表观遗传时钟。在随后的几年里，大量基于 DNA 甲基化的表观遗传时钟被开发出来，这些表观遗传时钟大多数都采用类似的方法，选择与年龄相关的高甲基化和低甲基化的关键 CpG 位点，将关键 CpG 位点数据输入线性模型中进行加权。结果得到一个包含每个关键 CpG 位点权重的方程，由此方程可以基于给定数据样本中这些关键 CpG 位点的 DNA 甲基化 beta 值来估计实际年龄。Weidner 等使用 3 个 CpG 位点开发用于血液年龄预测的表观遗传时钟[9]。McEwen 等使用 94 个 CpG 位点开发了用于 0 ~ 20 岁儿童年龄预测的儿童表观遗传时钟[10]，Horvath 等使用了 391 个 CpG 位点开发基于血液和皮肤的表观遗传时钟[11]等，都取得了不错的效果。

随着深度学习的发展，许多研究人员开始将深度学习技术用于 DNA 甲基化年龄预测领域。Thong 等使用 3 个 CpG 位点对比评估了线性回归模型和人工神经网络模型，证明了人工神经网络模型比线性回归模型具有更高的年龄预测准确性[12]。Li 等提出使用 Correlation Pre-Filtered Neural Network（CPFNN）进行年龄预测，发现适当加权与预测结果高度相关的特征是提高预测准确性的一个关键因素[13]。Camillo 等参考 Galkin 等基于血液样本训练的深度神经网络模型 DeepMAge[14]，提出一种使用深度神经网络的模型 AltumAge[15]。相比使用线性回归弹性网络的方法，该模型不仅提高了 DNA 甲基化年龄预测精度，而且在有关 CpG 位点之间的相关度问题上进行了一些讨论。

值得关注的是，所有的表观遗传时钟都选取了一定数量的 CpG 位点，但是这些 CpG 位点并不完全重合，每种不同组合的 CpG 位点都可以较准确地预测年龄。表观遗传时钟变得越来越多样化，同时也证明了它们在预测年龄方面是准确的，为衰老领域的研究提供了可靠的工具，这些工具对于研究衰老问题和研究与年龄有关的疾病具有重要的作用和意义。

2 材料与方法

2.1 数据可用性

数据集来源于公开可用的 Gene Expression Omnibus（GEO）数据库。共使用了来自 35 项研究的 DNA 甲基化数据集，包含 7 962 个样本，均来源于 Illumina Infinium HumanMethylation27（27K）和 Illumina Infinium HumanMethylation450（450K）平台。其中，训练数据中包含 27 项研究的数据集共 5 652 个不同组织的样本。其他 8 项研究的数据集共 2 310 个不同组织的样本作为独立验证集测试使用。所有数据集均可通过相应的 GEO 编号搜索到。

所有选定数据均是下载 GEO 数据库中经过处理的矩阵数据（Series Matrix File）。由于 GEO 数据库中部分数据存在缺失和格式不统一的情况，因此，首先，对下载的数据集中年龄缺失的样本进行去除，保证所有样本均有真实年龄值，以供后续使用。其次，对数据集表达矩阵具有缺失值的数据使用 methyLImp[16]进行填充。methyLImp 是一个基于线性回归的 DNA 甲基化数据的缺失值估计方法，该方法的基本原理在于观察到甲基化水平显示出高度的样本间相关性，通过相关性对缺失值进行补全。最后，从对数据集中提取我们所需的 21 368 个 CpG 位点的 DNA 甲基化数据。所用训练数据集如表 1 所示。

表 1　用于模型训练及测试的数据集

Tab. 1　Data set for model training and testing

ID	Platform	Tissue	Samples
GSE78874	450K	Saliva	259
GSE92767	450K	Saliva	54
GSE99029	450K	Saliva	57
GSE138279	450K	Saliva	65
GSE34035	27K	Saliva	197
GSE28746	27K	Saliva	84
GSE101961	450K	Breast	121
GSE72773	450K	Whole blood	310
GSE20236	27K	Whole blood	93
GSE72775	450K	Whole blood	335
GSE61496	450K	Whole blood	310
GSE87571	450K	Whole blood	727
GSE58045	27K	Whole blood	172
GSE56105	450K	Lymphocyte	614
GSE94876	450K	Buccal	120
GSE137688	450K	Buccal	250
GSE48988	27K	Colon	178
GSE51954	450K	Dermis, Epidermis	74
GSE36064	450K	Leukocyte	78
GSE27097	27K	Leukocyte	398
GSE137495	450K	Buccal	145
GSE137884	450K	Buccal	89
GSE137903	450K	Buccal	254
GSE137894	450K	Buccal	98
GSE94734	450K	Buccal	177
GSE137502	450K	Buccal	179
GSE56581	450K	T cell	214

2.2　MLPAge 模型设计

虽然近年来的表观遗传时钟在年龄预测方面取得了巨大的成就，但仍然存在两个问题。第一，不同的表观遗传时钟所选取的 CpG 位点极少重叠，是否选取越多的 CpG 位点就会有越好的结果？第二，绝大多数表观遗传时钟在选择 CpG 位点时，只考虑了与年龄具有线性关系的 CpG 位点，是否需要考虑更多与年龄具有非线性关系的 CpG 位点？

基于这两个问题，我们尝试实验更多的 CpG 位点数目，并且为了方便与其他方面进行对比实验，我们选择了 27K 和 450K 两种阵列数据。标准的 27K 和 450K 阵列数据中 CpG 位点数分别为 27 578 和 485 577，但 450K 中 CpG 位点并没有完全涵盖 27K 中的所有 CpG 位点。因此，为了使模型能够应用到两种阵列数据上，本文选取了 27K 和 450K 阵列数据共有部分的 CpG 位点数据，共 21 368 个 CpG 位点，作为本文模型的输入。

我们开发了 MLPAge，这是一个基于多层感知机的深度学习模型，使用来自 Illumina 27K 和 Illumina 450K 中重合的 21 368 个 CpG 位点的 beta 值进行训练和年龄预测。MLPAge 的模型设计图如图 1 所示。我们将年龄预测作为回归任务执行，其中模型将 21 368 个 CpG 位点的 DNA 甲基化 beta 值作为输入，然后输出连续的年龄值。在模型中，输入信息通过 4 个隐藏层进行处理，每个隐藏层都有 32 个节点。最后一个隐藏层节点的值组合成一个节点，输出年龄值。

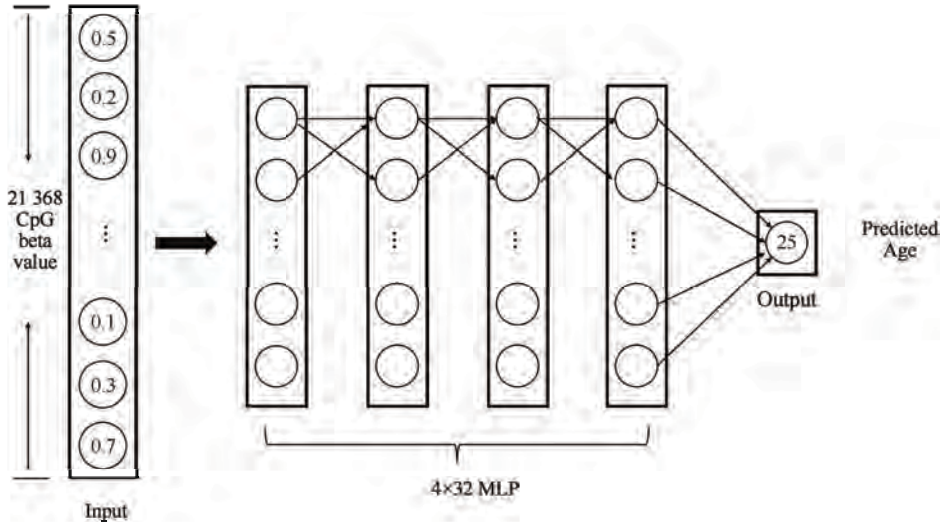

图 1　MLPAge 的模型设计图

Fig. 1　Model design of MLPAge

2.3　模型测试

首先，对所有 DNA 甲基化数据集进行导入，提取模型需要的 21 368 个 CpG 位点的 DNA 甲基化 beta 值，合并成为整个训练数据。由于 DNA 甲基化数据具有组织特异性，并且不同数据集的组织不同，需要将训练数据进行随机排序，将不同组织数据进行随机混合，防止组织特异性对实验结果造成影响。然后，对处理好的训练数据，按照 9∶1 的比例划分训练数据和测试数据。由于 DNA 甲基化数据本身为 0 ~ 1，因此并未对训练数据进行归一化处理。

训练数据以样本为批次输入模型中进行训练，输入数据为 $1 \times 21\,368$ 规格的一维数据。将输入数据依次放到隐藏层中进行处理，每个隐藏层都包含 32 个节点，前一个隐藏层处理后的结果都作为下一个隐藏层的输入，经过 3 个隐藏层的处理，最后一个隐藏层将数据合成一个数值输出。在模型训练迭代期间，每次送入的 BatchSize 均设置为 256。同时，考虑到 DNA 甲基化数据与年龄并非单纯的线性关系，因此每个隐藏层均引入 LeakyReLU()激活函数。为防止输入数据经过隐藏层处理之后产生数据分布改变，于是在每个隐藏层均添加 BatchNorm1d()函数对数据进行归一化处理。由于训练数据特征过大，模型很容易出现过拟合，特地在每个隐藏层中添加 Dropout()函数进行随机丢弃神经元。

3　结果分析

3.1　评价指标

为评价 MLPAge 预测年龄的准确度，采用相关系数（R^2）、平均绝对误差（MAE）、中位绝对误差（MAD）这 3 个评价指标。其中 MAE 如式（1）所示：

$$\mathrm{MAE} = \frac{1}{m}\sum_{i=1}^{m}\left(\mathrm{predAge}_i - \mathrm{trueAge}_i\right) \qquad (1)$$

MAD 如式（2）所示：

$$\mathrm{MAD} = \mathrm{median}\left(\left|\mathrm{predAge} - \mathrm{trueAge}\right|\right) \qquad (2)$$

式中，predAge 表示所有测试数据的预测年龄，trueAge 表示所有测试数据的实际年龄。

该模型在训练集和测试集上的评价指标如表 2 所示。综合其他文献常用的评价指标，最后将 MAE 和 MAD 作为本篇用于衡量预测结果准确度的评价指标。其中 R^2 仅用来展示实际年龄与预测年龄之间的相关性，其他两个指标用来评价预测年龄的准确度。

表 2　MLPAge 在训练集和测试集上的 3 个指标

Tab. 2　Three evaluationmetrics for the MLPAge onthe training andtest sets

数据类型	R^2	MAE	MAD
训练集	0.92	3.83	3.07
测试集	0.80	6.77	4.52

3.2　MLPAge 预测结果

MLPAge 测试结果可视化如图 2 所示。图 2（a）

是所有训练集样本的预测结果（年龄相关性 R^2=0.92，中位绝对误差 MAD=3.07 年），横轴为样本的实际年龄，纵轴为预测年龄。能够看出 MLPAge 在大部分数据集上表现效果极佳，只在部分样本的预测中存在误差较大的情况。图 2（b）是所有测试集样本的预测结果。实际年龄与预测年龄在图中均呈现较高的年龄相关性，测试集的 MAD=4.52 年。综合来看，样本模型预测准确度较高，并且在各个数据集中预测效果相对稳定。

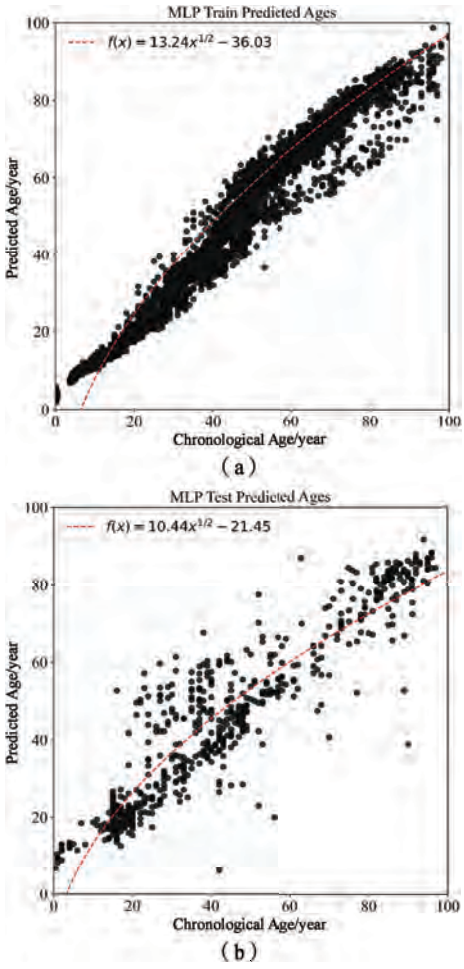

图 2　模型训练集与测试集结果可视化

Fig. 2　Visualization of model training set results and test set results

（a）模型训练集结果可视化；（b）模型测试集结果可视化

(a) Visualization of model training set results;

(b) Visualization of model test set results

3.3　与 Horvath 353CpG 时钟进行比较

为验证 MLPAge 预测的准确性，参考其他表观遗传时钟中使用频率较高的数据集，并综合考虑 Horvath 353CpG 时钟适用的数据阵列，收集了 8 个独立数据集，其中 27K 阵列和 450K 阵列数据集各 4 个，共 2 310 个样本。将 MLPAge 与当前该领域的标准 Horvath 353 CpG 时钟进行测试对比。对比结果如表 3 所示。

表 3　MLPAge 与 Horvath 353 CpG 时钟的对比结果

（红色为结果更优）

Tab. 3　Comparison of MLPAge and Horvath 353 CpG clock in results (better results in red)

Dataset	Platform	MAE MLPAge	MAD MLPAge	MAE Horvath	MAD Horvath
GSE25892	27K	2.09	2.06	2.94	2.91
GSE19711	27K	6.18	4.28	6.27	4.82
GSE111223	450K	6.57	6.55	10.50	9.73
GSE77136	450K	11.25	10.65	12.1	11.0
GSE72338	450K	5.28	4.16	4.17	3.3
GSE36194	27K	6.19	4.46	9.51	7.18
GSE41169	450K	3.75	2.77	2.95	2.13
GSE15745	27K	3.80	3.62	6.12	3.86

总体来看，MLPAge 在所使用的大部分独立数据集中的表现优于 Horvath 353 CpG 时钟。在所有 4 个 27K 阵列数据集中，MLPAge 的测试结果均优于 Horvath 353 CpG 时钟。由于 27K 阵列共包含 27 578 个 CpG 位点，而 MLPAge 选取了其中的 21 368 个，相较于 Horvath 353 CpG 时钟选取的 353 个 CpG 位点，在 27K 阵列数据集中占比更大，能够更加充分地提取数据的特征，在测试中的表现优于 Horvath 353 CpG 时钟。而从表 3 来看，MLPAge 和 Horvath 353 CpG 时钟在 4 个 450K 阵列数据集中的表现并不是很突出，原因在于 450K 阵列数据集共包含 485 577 个 CpG 位点，不管是 MLPAge 提取的 21 368 个 CpG 位点还是 Horvath 353 CpG 时钟提取的 353 个 CpG 位点，均占比不大，MLPAge 和 Horvath CpG 时钟在学习过程中对数据的学习不够充分，导致了部分结果偏低。

MLPAge 在 GSE72338 和 GSE41169 数据集中表现稍差，与 Horvath 353 CpG 时钟的结果相比偏高。分析原因发现，GSE41169 数据集中部分样本为精神分裂症样本，而 DNA 甲基化易受部分疾病影响，疾病会导致 DNA 甲基化水平出现异常，从而导致预测结果不准确。由于 MLPAge 训练时所选数据大多数为健康样本，训练数据集中相应疾病的样本数量较少，导致模型对该疾病数据特征的学习不足，进而误差过大，因此在测试时表现不佳。GSE72338 数据集中包含部分成纤维细胞，而训练样本中并不包含该种组织，模型对该组织的特征并不敏感，导致预测结果不佳。

同时，从结果看，GSE77136 数据集在 MLPAge 和 Horvath 353 CpG 时钟中的结果均不理想，分析发现，该数据为体外培养人类成纤维细胞中的 DNA 甲基化数据，体外培养人类成纤维细胞与人体正常细胞存在一定差异，所以该结果异常也在情理之中。

3.4 实验平台

MLPAge 使用深度学习框架 Pytorch 构建网络模型和用于数据预处理的 sklearn 库。Horvath 353 CpG 时钟采用 Horvath 的论文提供的 R 语言代码实现，并且保障代码测试结果与原文结果一致。实验软件环境为 Ubuntu 22.04, Python 3.8，Pytorch 1.13.0, sklearn 1.0.2，R 语言环境版本 R 4.1.1；实验硬件环境为阿里云服务器，内存 32 GB，GPU NVIDIA V100，CPU 8 核。

4 结论

准确的年龄预测方法能够为衰老领域的研究提供一个定量的生物年龄测量方法。本文探讨了基于深度学习的多组织 DNA 甲基化年龄预测方法的可用性、准确性，以及相比其他方法的优势和意义。虽然在人体衰老预测方法中，使用弹性网络方法已经预测得非常准确，但其对 CpG 位点之间相关关系的探究并不详细。深度学习模型在准确度上不仅能够略胜于弹性网络模型，而且更有助于研究 CpG 位点之间的相关关系。

但是，通过对比试验结果发现，MLPAge 在预测训练样本中不存在的组织和疾病数据时，仍然存在部分误差较大的情况。要想真正对多组织 DNA 甲基化年龄进行预测，需要进一步提高训练样本中不同组织样本的数量和比例，更平衡、更多样的训练数据会使模型在泛组织上的预测结果更加准确。同时，还需要考虑数据集中样本的疾病信息，不同疾病对 DNA 甲基化的影响各不相同，有些疾病会严重影响 DNA 甲基化水平，有些疾病则不会。因此，在处理训练数据时，应更多地关注数据集的疾病样本，进一步探究疾病与 DNA 甲基化的关系。最后，准确的表观遗传时钟能够通过预测年龄与真实年龄的对比，帮助临床医生判断部分与年龄相关的疾病的发生，更好地帮助临床医生进行临床诊断，对精准医疗具有一定的意义。

MLPAge 仍然有进一步改进的可能。在未来的工作中，需要进一步探究模型中 CpG 位点的变化，以及 CpG 位点之间的相互联系。应该充分考虑 CpG 组合与年龄的关系，而不是单一地一次性输入所有的 CpG 位点。例如，来源于同一染色体的 CpG 位点应该具有更高的相关性。进一步探究年龄与不同 CpG 位点及不同 CpG 位点组合之间的相关性，从中寻找与年龄具有高度相关性的 CpG 位点，挖掘衰老与 DNA 甲基化更深层的联系，以帮助我们进一步揭示衰老的原因。

本文提出了一种基于多层感知机的预测模型，与之前的利用弹性网络建立的表观遗传时钟相比，该模型的学习能力更强，能够提高预测的准确度。从独立测试数据集的 MAD 中能够看出，MLPAge 在大多数独立测试数据集上比 Horvath 353 CpG 时钟预测得更准确。MLPAge 为表观遗传时钟和衰老领域提供了一种新的基于深度学习的解决方案，代表了对当前表观遗传时钟方法的性能改进和深度学习技术在表观遗传时钟领域的可行性，为该领域提供了一种新的可行的生物学见解。

参考文献

[1] HORVATH S, RAJ K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing[J]. **Nature Reviews Genetics**, 2018, 19(6): 371-384.

[2] SIMPSON D J, CHANDRA T. Epigenetic age prediction[J]. **Aging cell**, 2021, 20(9): e13452.

[3] MORGAN A E, DAVIES T J, MC AULEY M T. The role of DNA methylation in ageing and cancer[J]. **Proceedings of the Nutrition Society**, 2018, 77(4): 412-422.

[4] LI X, PLONER A, WANG Y, *et al*. Longitudinal trajectories, correlations and mortality associations of nine biological ages across 20-years follow-up[J]. **Elife**, 2020, 9: e51507.

[5] BENAYOUN B A, POLLINA E A, BRUNET A. Epigenetic regulation of ageing: linking environmental inputs to genomic stability[J]. **Nature reviews Molecular cell biology**, 2015, 16(10): 593-610.

[6] DEDEURWAERDER S, DEFRANCE M, CALONNE E, *et al*. Evaluation of the Infinium Methylation 450K technology[J]. **Epigenomics**, 2011, 3(6): 771-784.

[7] HORVATH S. DNA methylation age of human tissues and cell types[J]. **Genome biology**, 2013, 14(10): 1-20.

[8] HANNUM G, GUINNEY J, ZHAO L, *et al*. Genome-wide methylation profiles reveal quantitative views of human aging rates[J]. **Molecular cell**, 2013, 49(2): 359-367.

[9] WEIDNER C I, LIN Q, KOCH C M, *et al*. Aging of blood can be tracked by DNA methylation changes at just three CpG sites[J].

Genome Biology, 2014, 15: 1-12.

[10] MCEWEN L M, O'DONNELL K J, MCGILL M G, *et al*. The PedBE clock accurately estimates DNA methylation age in pediatric buccal cells[J]. **Proceedings of the National Academy of Sciences**, 2020, 117(38): 23329-23335.

[11] HORVATH S, OSHIMA J, MARTIN G M, *et al*. Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and ex vivo studies[J]. **Aging (Albany NY)**, 2018, 10(7): 1758.

[12] THONG Z, TAN J Y Y, LOO E S, *et al*. Artificial neural network, predictor variables and sensitivity threshold for DNA methylation-based age prediction using blood samples[J]. **Scientific Reports**, 2021, 11(1): 1-12.

[13] LI L, ZHANG C, LIU S, *et al*. Age prediction by DNA methylation in neural networks[J]. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, 2021, 19(3): 1393-1402.

[14] GALKIN F, MAMOSHINA P, KOCHETOV K, *et al*. DeepMAge: a methylation aging clock developed with deep learning[J]. **Aging and Disease**, 2021, 12(5): 1252.

[15] CAMILLO L P, LAPIERRE L R, SINGH R. A pan-tissue DNA-methylation epigenetic clock based on deep learning[J]. **Npj Aging**, 2022, 8(1): 4.

[16] DI LENA P, SALA C, PRODI A, *et al*. Missing value estimation methods for DNA methylation data[J]. **Bioinformatics**, 2019, 35(19): 3786-3793.