

MAA: 数据集无关的衰老时钟评估研究

吕东旭

温州大学计算机与人工智能学院, 浙江 温州

收稿日期: 2025年3月15日; 录用日期: 2025年4月14日; 发布日期: 2025年4月22日

摘要

近年来, 基于不同衰老生物标志物的生物年龄预测得到了广泛的研究, 虽然先前的工作可以从不同的衰老生物标志物中构建对应的生物年龄预测模型, 但是由于所使用的数据集和生物标志物的差异, 作为主要评估指标的绝对平均误差在不同数据集之间的表现也存在一定的差异。为了减少数据集差异对模型评估产生的不必要的影响, 本文提出了一种新的评估指标MAA (Mean Absolute Accuracy), 通过以特定情况下的绝对平均误差为基准, 建立近似分类问题中的准确度度量标准。本文使用当前主流模型在不同的子数据集上进行了实验, 并与传统评估指标进行比较, 显示本文提出的MAA具有良好的数据集无关性, 为跨数据集和跨衰老生物标志物评估生物年龄预测模型的性能提供了新的工具。

关键词

生物年龄, 回归效应, 模型评估

MAA: Dataset-Independent Aging Clock Evaluation Research

Dongxu Lyu

College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou Zhejiang

Received: Mar. 15th, 2025; accepted: Apr. 14th, 2025; published: Apr. 22nd, 2025

Abstract

In recent years, biological age prediction based on different aging biomarkers has been extensively studied. Although previous work has constructed corresponding biological age prediction models from various aging biomarkers, the mean absolute error, which is the primary evaluation metric, shows some variation across different datasets due to differences in the datasets and biomarkers used. To reduce the unnecessary impact of dataset differences on model evaluation, this paper proposes MAA (Mean Absolute Accuracy) a new evaluation metric. By using the mean absolute error in

specific situations as a baseline, it establishes an accuracy measure in the context of approximate classification problems. This paper conducts experiments using current mainstream models on different sub-datasets and compares them with traditional evaluation metrics. The results show that the proposed MAA metric has good dataset independence, providing a new tool for evaluating the performance of biological age prediction models across datasets and aging biomarkers.

Keywords

Biological Age, Regression Effect, Model Evaluation

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

随着人工智能与机器学习技术的迅猛发展，模型评估成为研究和应用中的一个关键环节。评估方法不仅影响模型的选择与优化，还直接关系到模型在实际场景中的表现。生物年龄预测是一个不断发展但具有挑战性的研究领域，然而随着衰老时钟的快速发展，传统的评估指标可能无法全面反映模型在实际应用中的效果，作为主要评估指标的 MAE 在不同的衰老时钟之间可以达到 3~5 倍的差距，想要准确比较不同衰老时钟之间的真实性能比较困难。为了促进不同衰老时钟之间的比较，需要一种新的生物年龄预测模型的评估方法，并为不同衰老生物标志物提供更加精准的评估依据。

因此，建立一种可以实现跨数据集比较模型年龄预测效果的新的评估方法将更有利于不同衰老时钟的比较和后续关于不同衰老生物标志物的研究。本文提出了一个全新的生物年龄预测模型的评估方法 MAA，它能够克服传统指标 MAE 受数据集年龄区间影响的问题，对不同衰老时钟实现数据集无关的比较。这一新方法不仅能够提升模型评估的准确性，还能帮助研究人员和工程师更加高效地优化模型性能。

2. 相关工作

在早期的研究中，研究者通常使用机器学习模型结合批量组学数据来构建对应的衰老时钟。其中，DNA 甲基化时钟是最早和最广泛使用的衰老时钟之一。早在 2011 年，Bocklandt 等人就使用基于阵列的 DNA 甲基化数据和 Lasso 线性回归模型准确预测了人类年龄(MAE = 5.2) [1]。之后又发展出四种广泛使用的 DNA 甲基化衰老时钟：Horvath 时钟[2]，Hannum 时钟[3]，Levine 时钟[4]和 GrimAge 时钟[5]。

随着 DNA 甲基化时钟的成功，基于其他衰老生物标志物构建的衰老时钟也相继被提出。2015 年，Peters 等人基于 14,983 个个体全血的基因表达数据，使用多元线性回归构建了第一个转录组衰老时钟 (MAE = 7.77) [6]。同年，第一个基于 3D 面部图像的年龄预测模型被提出，其 MAE 为 6 年[7]。2017 年，Cole 等人使用 3D-CNN 网络在大脑 MRI 数据上生成了“大脑预测年龄”，MAE 为 4.16 年，这比传统的机器学习方法更好[8]。2018 年，Tanaka 等人利用血浆蛋白质组整体的弹性网络回归模型创建了一个衰老时钟[9]。2019 年，Lehallier 等人使用 lasso 模型和更多的血浆蛋白(n = 373) (PCC = 0.97)实现了更准确的年龄预测[10]。2021 年，Trapp 等人率先开发了第一个表观遗传单细胞衰老时钟，命名为 scAge [11]。Ahadi 等人在 2023 年建立了更准确的视网膜衰老时钟，并通过 GWAS 命中识别与几种年龄相关功能相关的候选基因来加强，从而加强了视网膜衰老时钟的生物学基础[12]。这些衰老时钟基于细胞、器官、有机体水平等不同层次和不同方面的衰老生物标志物来预测生物年龄，但由于这些数据较为敏感，因此用于预测

生物年龄的数据集在伦理隐私等方面有较高的要求，想要比较不同衰老时钟的性能需要一种数据集无关的评估指标。而在预测方法上，从传统的机器学习方法到近年来快速发展的深度学习方法都有涉及，预测方法的不同和衰老生物标志物的来源的复杂性，对衰老时钟的评估提出了更高的要求。

在生物年龄预测领域不断发展的过程中，衰老生物标志物不断增多，对应的数据集也不断增加，越来越多的研究者逐渐意识到单一的 MAE 评估指标无法准确衡量模型的真实性能。于是，后续研究中大多会使用多个指标来综合评估模型的性能，如：皮尔逊相关系数(Pearson Correlation Coefficient, PCC)，RMSE (Root Mean Square Error)等。这些指标的增加使得对当前模型的综合评估更为全面，但是鉴于不同年龄预测模型是基于不同的数据集来构建的，因此在跨数据集的模型评估中作为主要指标的 MAE 仍然存在较大差异。与他们不同的是，本文并不是通过增加传统指标来完善模型评估的，而是通过设计一个新的数据集无关的评价指标 MAA 来实现，MAA 是以特定情况下的绝对平均误差为基准建立的近似分类问题中的准确度度量标准，它使得我们可以在同一度量范围上实现跨数据集的生物年龄预测模型的性能评估。

3. 方法

本文的主要目标是通过研究不同年龄预测数据集之间的主要差异和当前主流模型的年龄预测表现来设计一种新的生物年龄预测模型的评价指标，它可以跨越不同数据集、不同衰老生物标志物来衡量不同衰老时钟的性能。同时为了确认 MAA 的有效性，我们将使用主流的 CNN (Convolutional Neural Network) 架构的网络模型和 ViT (Vision Transformer)架构的网络模型来进行相同的实验。下面我们将从年龄区间、回归效应和评估指标三个方面进行方法介绍。

3.1. 年龄区间

我们收集一些生物年龄预测领域的数据集相关信息，如表 1 所示。通过对这些数据集的对比分析可以发现这些数据集的构成存在明显不同，对应的模型的预测性能也存在显明差异。

Table 1. Comparison between different Datasets
表 1. 不同数据集之间的比较

Reference	Dataset	Age Range	Age Span	Samples	Data Type	MAE	PCC
Han <i>et al.</i> [13]	HCP, CamCAN, IXI	18~88	70	2281	T1 MRI	2.75	0.43
Ly <i>et al.</i> [14]	ADNI, IXI, OASIS-3	20~85	65	1248	T1 MRI	4.65	0.6
Beck <i>et al.</i> [15]	TOP, StrokeMRI	18~95	77	702	dMRI	6.99	0.85
Xifra-Porxas <i>et al.</i> [16]	CamCAN	18~88	70	613	T1 MRI, MEG	4.88	0.94
Tesli <i>et al.</i> [17]	TOP, StrokeMRI	12~92	80	586	T1 MRI	6.6	0.68
Ballester <i>et al.</i> [18]	COBRE, MCIC, etc	16~65	49	471	fMRI	5.61	0.8

由表 1 可知，不同论文中使用的用于生物年龄预测的数据集中，其首要差异就是使用的数据类型不同，除此之外，这些数据集之间的主要差异就在于不同数据集的样本所对应的年龄区间的不同。而且即使是在使用相同衰老生物标志物的数据集中，由于不同论文中使用的处理方法不同也会造成年龄区间的变化。因此，研究不同年龄区间对最终结果的影响是十分有必要的。为了比较不同年龄区间的实际影响，并排除不同数据类型的影响，我们会将同一个数据集按不同年龄区间划分为对应的子数据集进行对应的实验来观察结果的变化趋势。

3.2. 回归效应

在统计学中,回归效应表现为数据点的分布趋势。当数据点偏离其均值时,会有一种趋势使它们逐渐回到均值附近。在年龄预测的回归任务中同样受其影响,并且经过我们的实验测试发现,这种回归效应的影响对不同的神经网络模型而言,普遍存在。具体表现为:当样本的实际年龄偏低时,往往预测年龄相对真实年龄偏高;当样本的实际年龄偏高时,往往预测年龄相对真实年龄偏低。同时,在回归效应的影响下,年龄的预测几乎总是被限制在训练年龄的范围之内,对处于训练集年龄区间两端的年龄总是会表现出特定的倾向。

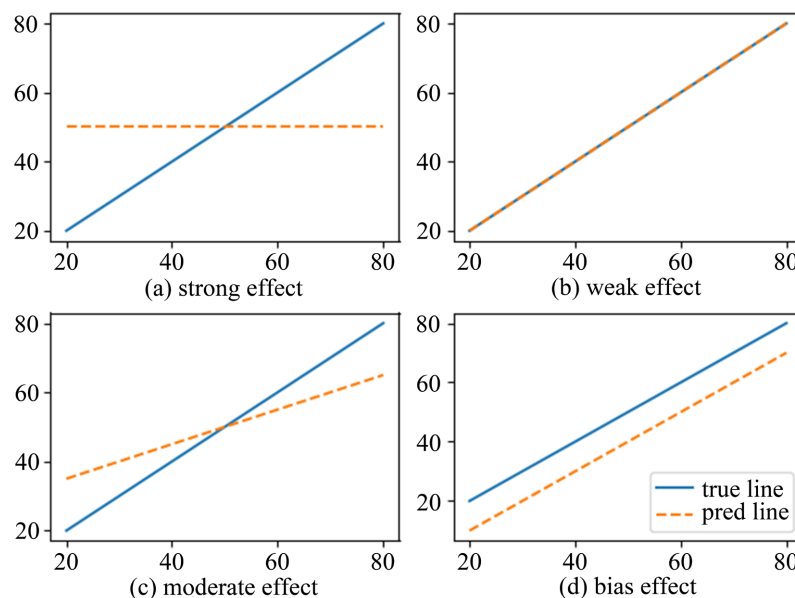


Figure 1. The illustrative diagram of regression effect

图 1. 回归效应影响示意图

如图 1 所示,受回归效应影响程度的不同,预测结果进行线性拟合后大致可以分为四种情况,其中蓝色实线表示理想情况下预测结果线性拟合后的直线,黄色虚线表示实际情况下预测结果线性拟合后的直线。其中,图 1(a)表示回归效应对预测结果产生了较强的影响的情况,此时预测值始终处于均值附近,模型完全受到回归效应的支配,无法进行有效训练,也无法完成年龄预测的任务;图 1(b)表示回归效应对预测结果产生了较弱的影响的情况,此时预测值始终处于实际值附近,模型完全不受回归效应的影响,可以进行有效训练,也可以完成年龄预测的任务;图 1(c)表示回归效应对预测结果产生了一定程度的影响的情况,此时预测值和实际值有一定的正相关性,但程度有所不同,模型可以得到有效训练,但会一定程度上受到回归效应的影响,回归效应被抑制的情况下可以完成年龄预测任务;图 1(d)表示预测值并未向均值回归的情况,此时预测值可能大部分都高于或者低于实际值,模型得到了一定程度的训练,但是预测值出现了总体性的偏移,这种偏移带来的误差会导致模型整体误差增大,在偏移带来的误差不大且回归效应的影响不深时,模型仍然可以得到有效训练,也可以完成年龄预测的任务。

在实际情况下,子图 1(a)和子图 1(b)所代表的情况并不常见,子图 1(c)和子图 1(d)所代表的情况则是目前常见的情况。但是如果一个模型如果能够形成有效的年龄预测,则它的情况应该处于子图 1(a)与子图 1(b)之间。于是可以利用子图 1(a)的情况为基准来计算得到 MAA。由此可见,如果能够对回归效应的影响程度和偏移误差进行数据集无关的量化比较,就可以实现跨数据集的模型性能评估。

3.3. 评估指标

在本文中,我们将使用 MAE, PCC 和 MAA 使用评估指标对每一次实验的结果进行评估。

MAE (Mean Absolute Error)就是平均绝对误差,它作为评估回归模型的一种指标,是一个相对直观的统计学指标,在回归分析中作为一个基础的评价方法被广泛采用。MAE 具有与原始数据相同的单位,直接表示预测值与真实值的平均差距,因而易于理解和解释。将 MAE 记为 ε 。它的计算公式如下:

$$\varepsilon = \text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (1)$$

其中, y_i 是实际值, $\hat{y}_i$ 是预测值, n 是样本数量。

PCC (Pearson Correlation Coefficient)是皮尔逊相关系数,它是衡量两个变量之间线性相关性的一个统计量。PCC 的值介于-1 和 1 之间,表示两个变量之间的线性关系强度和方向。需要注意的是,PCC 对异常值非常敏感,少量的异常数据点可能会显著影响结果。它的计算公式如下:

$$\text{PCC} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (2)$$

其中, x_i 是实际值, $\bar{x}$ 是实际值的均值, y_i 是预测值, $\bar{y}$ 是预测值的均值, n 是样本数量。

MAA 以基准误差进行计算的,因此我们需要先计算其基准误差的大小,按上一小节确定的基准情况,该情况下计算得到的 MAE 为基准误差,记为 ∂ ,于是它的计算公式如下:

$$\partial = \frac{1}{n} \sum_{i=1}^n |y_i - \bar{y}| \quad (3)$$

其中, y_i 是实际值, $\bar{y}$ 是实际值的均值, n 是样本数量。

于是,MAA 的计算公式可以表达如下:

$$\text{MAA} = \frac{\partial - \varepsilon}{\partial} \times 100\% \quad (4)$$

由公式(4)可知,MAA 反映的是 MAE 相对于基准情况下 MAE 的差异,也是两者差值占基准误差 ∂ 的比例。于是,MAA 形成了近似分类问题中准确度的评估形式,但是二者在原理和表现上仍有不同,比如 MAA 的值几乎永远不可能为 1,只能趋近于 1;且 MAA 可以为负,此时说明模型的预测出现了极大的偏差,几乎处于不可用的状态,也无法完成年龄预测的任务。所以可以规定当基准误差 ∂ 小于实际误差 ε 时,MAA 为 0。显然,当 MAA 越接近 0 时,模型的预测误差越大,预测效果相对较差;当 MAA 越接近 1 时,模型的预测误差越小,预测效果相对较好。

4. 实验

4.1. 数据集

本实验使用了一个公开可用的二维人脸数据集 AgeDB 来进行训练和测试。该数据集是一个专门收集了不同年龄段人脸图像的数据集,适用于人脸识别、年龄估计等相关领域的研究和测试。该数据集包含了多样化的年龄分布,能够帮助研究人员和开发者评估和改进人脸识别算法的性能[19]。我们将 AgeDB 中的一张图像视为一个样本,并对其基本信息进行了统计分析和数据处理得到最终数据。

在原始的 AgeDB 中共包含 16,488 个样本,年龄范围在 1~101 之间,平均年龄为 46.3 岁。其中,男

性样本 9788 个，年龄范围在 1~97 之间，平均年龄为 48.1 岁；女性样本 6700 个，年龄范围在 3~101 之间，平均年龄 43.5 岁。为了便于进行训练集和测试集的划分，我们将其中样本数量为一样本去除，对应的年龄分别为 1、97、99 和 101，最后得到可用于训练的完整数据集，如下表所示：

Table 2. AgeDB gender and age information

表 2. AgeDB 性别和年龄信息

Dataset	Gender	Samples	Age Range	Mean Age	Median Age
AgeDB	Female	6698	3~100	43.5119	40
	Male	9786	3~95	48.1314	47
	All	16,484	3~100	46.2544	44

从表 2 可知，最终用于模型训练和测试的 AgeDB 数据集的总体年龄区间在 3~100 之间，男性样本的年龄区间略小于女性，但男性样本的数量、平均年龄和中位数年龄均高于女性。总体的平均年龄为 46.2544，中位数年龄为 44 岁，相对于数据集的年龄区间而言总体年龄偏年轻化。

最后，由于 AgeDB 中的图像通道数量存在差异，因此先统一转换为 RGB 格式，再统一裁剪为 224×224 的大小，并在输入网络前进行随机的水平翻转和垂直翻转操作来增加数据的多样性，从而提高模型的泛化能力，最后以分层抽样的方法划分训练集和测试集，其中训练集占比 80%，测试集占比 20%。

4.2. 训练测试

我们将使用主流的深度学习模型在不同年龄区间的 AgeDB 数据集中进行实验。首先将所选定的模型在同等条件下，用 AgeDB 数据集进行了训练和测试。然后通过不断缩减 AgeDB 的年龄区间来选定样本，并对这些选定的样本在同样的条件下进行实验。最后对不同模型架构在不同年龄区间上的表现进行定性分析和定量评估。

本实验使用统一的运行环境，在同一台 Linux 服务器上运行，该服务器配备了一个 Intel Xeon Gold 6253L CPU，内存总容量为 251GB，2 个 NVIDIA GeForce RTX 3090 GPU，每个 GPU 具有 24GB 的显存，总共使用了 1 个 GPU 进行训练。所有的深度学习训练和推理过程都在 Ubuntu 22.04 系统上进行，我们使用的深度学习框架为 PyTorch 2.2.1，对应的 Python 版本为 3.11.8，并使用 550.67 版本的 NVIDIA 驱动，通过 CUDA 12.4 在 NVIDIA GeForce RTX 3090 GPU 上进行加速计算。

我们对所有模型都采用不同的超参数进行训练，这些超参数如表 3 所示：

Table 3. Hyperparameters for model training

表 3. 模型训练的超参数

Hyperparameters	Value
Optimizer	Adam
Learning Rate	$1e^{-2}$
LR Scheduler	ReduceLROnPlateau
Patience	15
Epochs	125
Batch Size	8

AgeDB 的年龄区间在 0~100 岁，我们通过不断缩小其年龄区间来构建年龄范围不同的子数据集，其年龄区间依次缩小为 10~90，20~80，30~70，40~60，并在这些子数据集上按相同的标准进行实验。

4.3. 定性评估

我们通过各模型在不同年龄区间上的预测结果的散点图来直观地了解年龄区间的变化和回归效应对预测结果的影响之间的关系。下面所有的散点图都遵循以下几点规则：1) x 轴表示实际年龄，y 轴表示预测年龄。2) 蓝色实线表示理想情况下预测结果线性拟合后的直线，其斜率为 1。3) 黄色虚线表示实际情况中预测结果线性拟合后的直线，其斜率应在 0~1 之间。下面是不同模型在不同年龄区间的子数据集上实际年龄与预测年龄的散点图。

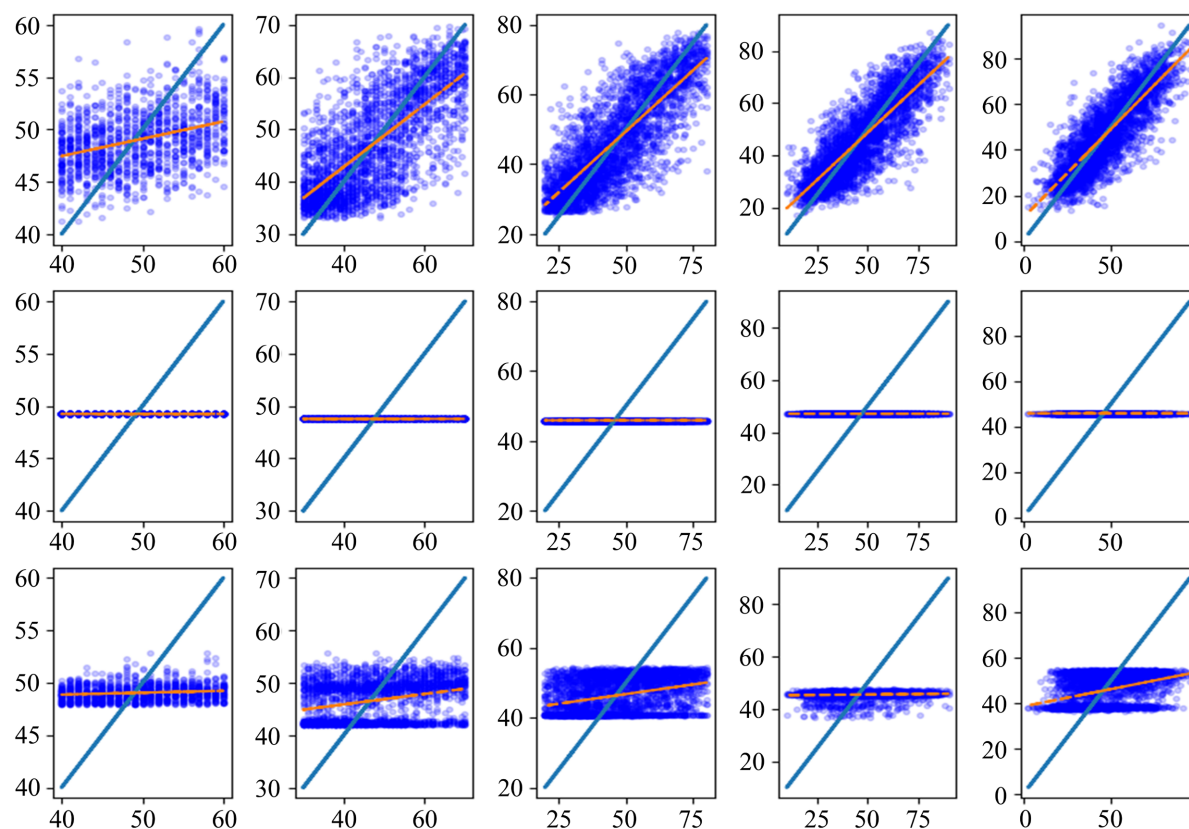


Figure 2. The scatter plot of ViT architecture model on AgeDB

图 2. ViT 架构模型在 AgeDB 上的散点图

如图 2 所示，每一列表示从 AgeDB 数据中选取不同年龄范围的样本来构成新的数据集用来训练和测试，从左到右分别表示的年龄区间为：40~60，30~70，20~80，10~90，0~100。每一行表示一个 ViT 架构的模型在不同年龄区间的数据集上训练和测试的表现，从上到下分别使用的模型为：FastViT，MaxViT，Visformer。从图像上看，MaxViT 和 Visformer 在每个年龄区间的子数据集上都没能形成有效训练，都接近图 1(a)的情况，完全被回归效应支配。而 FastViT 在 40~60 年龄区间的数据集上的训练效果不佳，但随着年龄区间的增大，回归效应逐渐得到抑制，越来越接近图 1(b)的情况，预测效果逐渐变好。

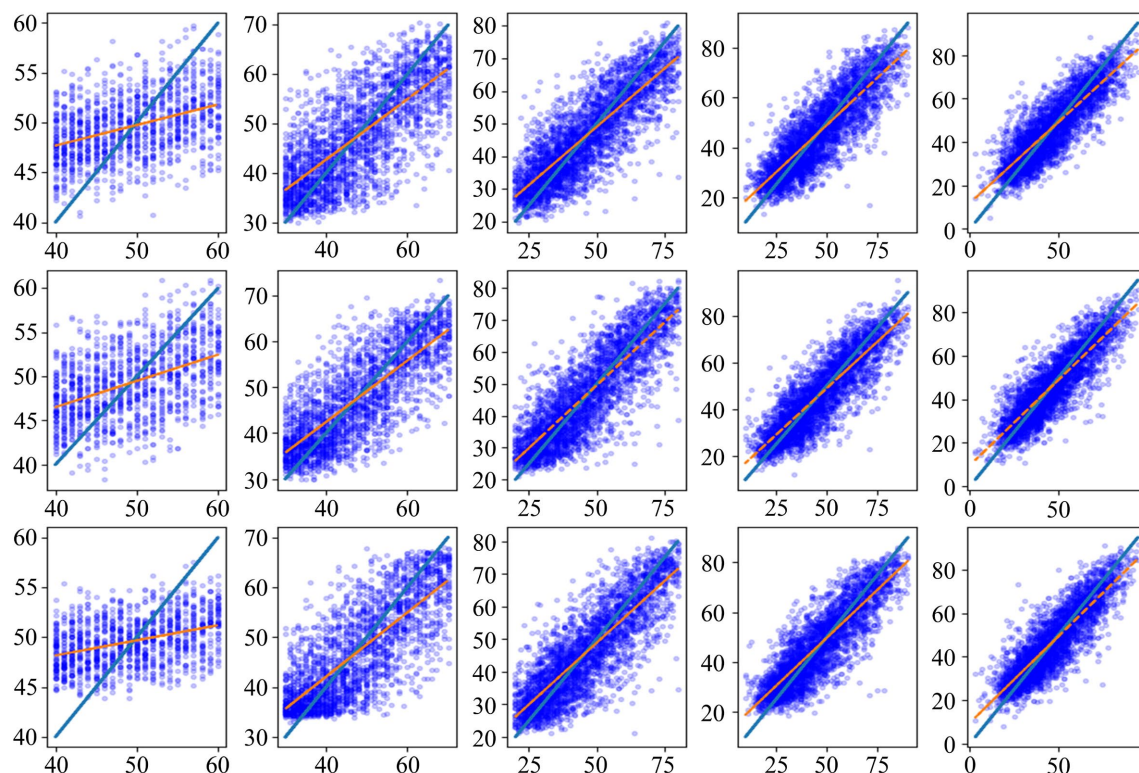


Figure 3. The scatter plot of CNN architecture model on AgeDB

图 3. CNN 架构模型在 AgeDB 上的散点图

如图 3 所示，每一列表示从 AgeDB 数据中选取不同年龄范围的样本来构成新的数据集用来训练和测试，从左到右分别表示的年龄区间为：40~60，30~70，20~80，10~90，0~100。每一行表示一个 CNN 架构的模型在不同年龄区间的数据集上训练和测试的表现，从上到下分别使用的模型为：ResNet，Xception，EfficientNet。三个模型在不同年龄区间的数据集上和 FastViT 有相同的表现，在 40~60 年龄区间的数据集上的训练效果都不佳，但随着年龄区间的增大，回归效应逐渐得到抑制，预测效果逐渐变好。

综上所述，虽然部分 ViT 架构的模型无法形成有效的年龄预测，但是在可以形成有效训练的多个模型上，年龄区间和回归效应的影响都有相似的表现：数据集的年龄区间越小，回归效应的影响程度越大；反之，年龄区间越大，回归效应的影响程度越小。

4.4. 定量分析

本小节中，我们将通过不同的评估指标来定量分析随着年龄区间变化的各评价指标的变化，以下的两张表格分别展示了 ViT 架构的模型和 CNN 架构的构型在不同年龄区间的数据集上的表现，结果如下所示。

表 4 中展示了 ViT 架构的三个模型评价指标的变化情况。虽然在上一小节的定性评估中我们已经知道 MaxViT 和 Visformer 并没有形成有效训练，但是三者的 MAE 都一致地表现为随着年龄区间的增大而增大的趋势，只有 Visformer 在 1~100 的年龄区间上的结果有所不同，但鉴于其和 10~90 时的 MAE 相差不大，且 Visformer 并未形成有效训练，所以我们仍然可以认为上述结论成立。接着，我们可以发现在形成有效训练的 FastViT 模型中，MAA 和 PCC 也都随着年龄区间的增大而增大，这表明模型的预测效果也随着年龄区间的增大而变好。但如果单以 MAE 来看的话，则是误差变大，预测效果变差，显然这与我们在上一小节中观察到的现象不符，而 MAA 的表现更接近实际情况。

Table 4. The performance metrics of ViT architecture on AgeDB
表 4. ViT 架构模型在 AgeDB 上的性能指标

Model	Age Range	MAE	MAA	PCC
FastViT	40~60	4.8139	0.0754	0.671
	30~70	6.2937	0.3618	0.8594
	20~80	7.076	0.4624	0.9058
	10~90	7.2355	0.4951	0.9184
	0~100	7.2654	0.4996	0.9217
MaxViT	40~60	5.2064	0	0.4745
	30~70	9.8677	-0.0006	0.5731
	20~80	13.1515	0.0009	0.6112
	10~90	14.4094	-0.0055	0.4902
	0~100	14.4891	0.002	0.5113
Visformer	40~60	5.1388	0.013	0.5711
	30~70	9.2168	0.0654	0.6485
	20~80	12.178	0.0749	0.6738
	10~90	14.2231	0.0075	0.5428
	0~100	13.0987	0.0977	0.6946

Table 5. The performance metrics of CNN architecture on AgeDB
表 5. CNN 架构模型在 AgeDB 上的性能指标

Model	Age Range	MAE	MAA	PCC
ResNet	40~60	4.6396	0.1089	0.6984
	30~70	6.1825	0.3731	0.8644
	20~80	6.6732	0.493	0.9164
	10~90	7.0737	0.5064	0.9247
	0~100	6.9149	0.5237	0.9277
Xception	40~60	4.5756	0.1212	0.7221
	30~70	5.6722	0.4248	0.8897
	20~80	6.4567	0.5095	0.923
	10~90	6.7469	0.5292	0.9311
	0~100	6.5582	0.5483	0.9352
EfficientNet	40~60	4.7217	0.0931	0.6864
	30~70	5.7269	0.4193	0.8862
	20~80	6.4975	0.5064	0.9199
	10~90	6.6388	0.5367	0.9321
	0~100	6.5535	0.5486	0.9355

表 5 中展示了 CNN 架构的三个模型评价指标的变化情况。总体而言，三个模型在 MAE, MAA 和 PCC 上都有相近的数值变化趋势，排除在 0~100 的年龄区间上各模型的 MAE 出现小幅下降之外，几乎都随着年龄区间的增大而增大。这和 FastViT 的表现十分接近，只是从数值来看 CNN 架构的模型略优一些，再结合图 3 可以看出，MAE 变大并不能直接反映模型性能变差，而 MAA 和 PCC 的增大更符合模型在图 3 中的表现，即随着年龄区间的增大，模型的预测效果逐渐变好，这与 MAA 和 PCC 的变化方向一致。

5. 结论

本文提出的 MAA 能够实现数据集无关的生物年龄预测模型的性能评估。我们的实验结果表明，传统上最广泛使用的 MAE 对跨数据集的比较有一定的困难，它会随着数据集所包含的年龄区间的不断变小而减小。而本文提出的 MAA 则可以克服这一问题，MAA 能够更客观地反映建立在不同数据集上的模型的性能，不会因为数据集年龄区间的改变而受到影响，为跨越不同的衰老生物标志物以比较不同衰老钟的性能提供了新的工具。

关于模型的评价指标是一个十分重要的研究内容，它的好坏直接决定了对模型的好坏的衡量，它对于不同数据集之间的包容性也决定了基于不同数据集的模型之间比较的可靠性。因此，对生物年龄预测模型性能的影响因素进行更深入的探索，以建立更全面、更系统、更通用的评估标准会是未来的研究方向。

参考文献

- [1] Bocklandt, S., Lin, W., Sehl, M.E., Sánchez, F.J., Sinsheimer, J.S., Horvath, S., *et al.* (2011) Epigenetic Predictor of Age. *PLOS ONE*, **6**, e14821. <https://doi.org/10.1371/journal.pone.0014821>
- [2] Horvath, S. (2013) DNA Methylation Age of Human Tissues and Cell Types. *Genome Biology*, **14**, Article No. 3156. <https://doi.org/10.1186/gb-2013-14-10-r115>
- [3] Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sada, S., *et al.* (2013) Genome-Wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Molecular Cell*, **49**, 359-367. <https://doi.org/10.1016/j.molcel.2012.10.016>
- [4] Levine, M.E., Lu, A.T., Quach, A., Chen, B.H., Assimes, T.L., Bandinelli, S., *et al.* (2018) An Epigenetic Biomarker of Aging for Lifespan and Healthspan. *Aging*, **10**, 573-591. <https://doi.org/10.18632/aging.101414>
- [5] Lu, A.T., Quach, A., Wilson, J.G., Reiner, A.P., Aviv, A., Raj, K., *et al.* (2019) DNA Methylation Grimace Strongly Predicts Lifespan and Healthspan. *Aging*, **11**, 303-327. <https://doi.org/10.18632/aging.101684>
- [6] Petears, M.J., Joehanes, R., Pilling, L.C., *et al.* (2015) The Transcriptional Landscape of Age in Human Peripheral Blood. *Nature Communications*, **6**, Article No. 8570.
- [7] Chen, W., Qian, W., Wu, G., Chen, W., Xian, B., Chen, X., *et al.* (2015) Three-Dimensional Human Facial Morphologies as Robust Aging Markers. *Cell Research*, **25**, 574-587. <https://doi.org/10.1038/cr.2015.36>
- [8] Cole, J.H., Poudel, R.P.K., Tsagkrasoulis, D., Caan, M.W.A., Steves, C., Spector, T.D., *et al.* (2017) Predicting Brain Age with Deep Learning from Raw Imaging Data Results in a Reliable and Heritable Biomarker. *NeuroImage*, **163**, 115-124. <https://doi.org/10.1016/j.neuroimage.2017.07.059>
- [9] Tanaka, T., Biancotto, A., Moaddel, R., Moore, A.Z., Gonzalez-Freire, M., Aon, M.A., *et al.* (2018) Plasma Proteomic Signature of Age in Healthy Humans. *Aging Cell*, **17**, e12799. <https://doi.org/10.1111/acer.12799>
- [10] Lehallier, B., Gate, D., Schaum, N., Nanasi, T., Lee, S.E., Yousef, H., *et al.* (2019) Undulating Changes in Human Plasma Proteome Profiles across the Lifespan. *Nature Medicine*, **25**, 1843-1850. <https://doi.org/10.1038/s41591-019-0673-2>
- [11] Trapp, A., Kerepesi, C. and Gladyshev, V.N. (2021) Profiling Epigenetic Age in Single Cells. *Nature Aging*, **1**, 1189-1201. <https://doi.org/10.1038/s43587-021-00134-3>
- [12] Ahadi, S., Wilson, K.A., Babenko, B., McLean, C.Y., Bryant, D., Pritchard, O., *et al.* (2023) Longitudinal Fundus Imaging and Its Genome-Wide Association Analysis Provide Evidence for a Human Retinal Aging Clock. *eLife*, **12**, e82364. <https://doi.org/10.7554/elife.82364>

-
- [13] Han, J., Kim, S.Y., Lee, J. and Lee, W.H. (2022) Brain Age Prediction: A Comparison between Machine Learning Models Using Brain Morphometric Data. *Sensors*, **22**, Article 8077. <https://doi.org/10.3390/s22208077>
 - [14] Ly, M., Yu, G.Z., Karim, H.T., Muppidi, N.R., Mizuno, A., Klunk, W.E., *et al.* (2020) Improving Brain Age Prediction Models: Incorporation of Amyloid Status in Alzheimer's Disease. *Neurobiology of Aging*, **87**, 44-48. <https://doi.org/10.1016/j.neurobiolaging.2019.11.005>
 - [15] Beck, D., de Lange, A.G., Maximov, I.I., Richard, G., Andreassen, O.A., Nordvik, J.E., *et al.* (2021) White Matter Microstructure across the Adult Lifespan: A Mixed Longitudinal and Cross-Sectional Study Using Advanced Diffusion Models and Brain-Age Prediction. *NeuroImage*, **224**, Article ID: 117441. <https://doi.org/10.1016/j.neuroimage.2020.117441>
 - [16] Xifra-Porxas, A., Ghosh, A., Mitsis, G.D. and Boudrias, M.H. (2019) Estimating Brain Age from Structural MRI and MEG Data: Insights from Dimensionality Reduction Techniques. *NeuroImage*, **231**, Article ID: 117822.
 - [17] Tesli, N., Bell, C., Hjell, G., Fischer-Vieler, T., I Maximov, I., Richard, G., *et al.* (2022) The Age of Violence: Mapping Brain Age in Psychosis and Psychopathy. *NeuroImage: Clinical*, **36**, Article ID: 103181. <https://doi.org/10.1016/j.nicl.2022.103181>
 - [18] Ballester, P.L., Suh, J.S., Ho, N.C.W., Liang, L., Hassel, S., Strother, S.C., *et al.* (2023) Gray Matter Volume Drives the Brain Age Gap in Schizophrenia: A SHAP Study. *Schizophrenia*, **9**, Article No. 3. <https://doi.org/10.1038/s41537-022-00330-z>
 - [19] Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I. and Zafeiriou, S. (2017) AgeDB: The First Manually Collected, In-The-Wild Age Database. 2017 *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, 21-26 July 2017, 1997-2005. <https://doi.org/10.1109/cvprw.2017.250>