2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) | 979-8-3503-1892-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/WACV57701.2024.00770

Ordinal Classification with Distance Regularization for Robust Brain Age Prediction

Jay Shah[1,2], Md Mahfuzur Rahman Siddiquee[1,2], Yi Su[2,3], Teresa Wu[1,2], Baoxin Li[1,2]

[1]Arizona State University, [2]ASU-Mayo Center for Innovative Imaging,
[3]Banner Alzheimer's Institute

Abstract

Age is one of the major known risk factors for Alzheimer's Disease (AD). Detecting AD early is crucial for effective treatment and preventing irreversible brain damage. Brain age, a measure derived from brain imaging reflecting structural changes due to aging, may have the potential to identify AD onset, assess disease risk, and plan targeted interventions. Deep learning-based regression techniques to predict brain age from magnetic resonance imaging (MRI) scans have shown great accuracy recently. However, these methods are subject to an inherent regression to the mean effect, which causes a systematic bias resulting in an overestimation of brain age in young subjects and underestimation in old subjects. This weakens the reliability of predicted brain age as a valid biomarker for downstream clinical applications. Here, we reformulate the brain age prediction task from regression to classification to address the issue of systematic bias. Recognizing the importance of preserving ordinal information from ages to understand aging trajectory and monitor aging longitudinally, we propose a novel ORdinal Distance Encoded Regularization (ORDER) loss that incorporates the order of age labels, enhancing the model's ability to capture age-related patterns. Extensive experiments and ablation studies demonstrate that this framework reduces systematic bias, outperforms state-of-art methods by statistically significant margins, and can better capture subtle differences between clinical groups in an independent AD dataset. Our implementation is publicly available at https://github.com/jaygshah/Robust-Brain-Age-Prediction.

1. Introduction

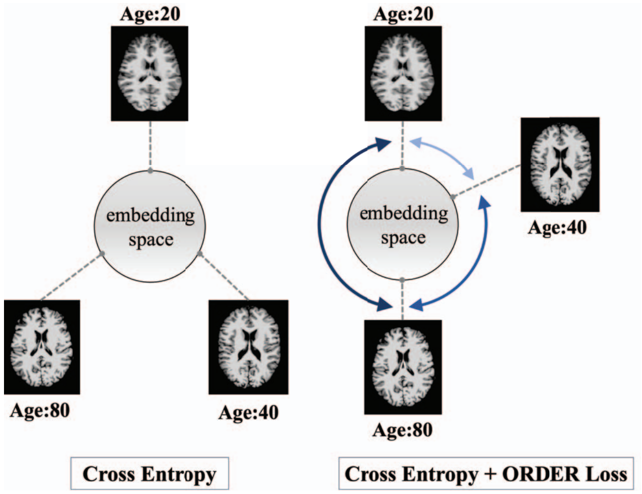


Figure 1. Standard cross-entropy vs. cross-entropy with ORDER loss: Cross entropy loss (left) encourages the model to learn high entropy feature representations where embeddings are spread out. However, it fails to capture ordinal information from labels. Our proposed ORDER loss with cross entropy (right, Eq. 5) preserves ordinality by spreading the features proportional to Manhattan distance between normalized features weighted by absolute age difference. The illustrated example (right) shows embedding space where learned representations of MRI scans with ages 20, 40, and 80 are distributed apart from one another, with distances proportional to absolute age differences.

Normal aging causes structural changes in the human brain across the adult lifespan, a major risk factor for the decline in physical health and cognitive ability [10]. Aging also exposes an individual to an increased risk of cancer [31] and various neurological disorders such as Parkinson's Disease [4], vascular dementia [20], mild cognitive impairment (MCI) and Alzheimer's Disease (AD) [15]. However, aging in humans is a complex and heterogeneous phenomenon. Even though each individual ages at the same rate chronologically, their biological age does not follow the same trajectory due to genetic factors, environmental influences, underlying neurological conditions, and other unknown factors [31]. Measuring this deviation from normal aging can allow a better understanding of associations between cognitive impairment and aging [14, 18] and identify patients at risk for clinical trials [10]. Hence, there is a growing interest in predicting biological age, most commonly derived from an individual's structural

2642-9381/24/$31.00 ©2024 IEEE
DOI 10.1109/WACV57701.2024.00770

MRI data. The difference between predicted biological age and chronological age, also known as Brain Age Gap Estimate (BrainAGE) [16], can be used to monitor accelerated or decelerated brain aging.

Measuring deviation from normal aging relies heavily on the base model's performance to predict normal aging, i.e., accurately predicting the biological age of healthy subjects. A model's performance on a healthy cohort is often assessed using mean absolute error (MAE), which calculates the mean of absolute BrainAGE. Existing deep learning-based regression approaches [23, 9, 21] have limited clinical applications because the models have reported MAE $4-5$ years in healthy cohorts, suggesting the lack of discriminative power to interrogate BrainAGEs of different clinical groups [18]. Moreover, a common challenge in brain prediction models is the issue of systematic bias [5, 27, 47], where there is an overestimation in the predicted age of young subjects and an underestimation of old subjects. If BrainAGE were to be used as a reliable imaging biomarker for measuring brain health, the effect of systematic bias is of concern. For example, Alzheimer's disease patients are often of age 50+, and age underestimates will impact early detection. Studies have investigated whether this bias is induced due to model or data selection used for training [27]. We argue that systematic bias is inherent to brain age prediction due to its formulation as regression analysis. This study has two primary objectives: (1) addressing systematic bias to enhance the robustness of brain age prediction and (2) enhancing the model's performance in predicting normal aging in healthy cohorts, thereby facilitating more accurate disease detection in downstream tasks.

Traditionally, brain age estimation is formulated as a regression task since the problem of interest is understanding which bio-signatures from imaging data have a statistically significant effect on age. More importantly, it is clinically relevant to study how these signatures change across different age groups and track their progression. To accomplish this, capturing ordinal information from the ground-truth age is critical; hence, regression is preferable. However, it is known that regression models suffer from systematic bias. To address this issue, we propose reformulating the task of brain age prediction as a multi-class classification. However, in classification, each class is treated independently of the other and hence cannot capture the ordinality of target labels [48]. To counter this, we propose a novel *ORdinal Distance Encoded Regularization* (ORDER) loss in conjunction with cross-entropy loss for multi-class ordinal classification. ORDER loss is calculated based on the Manhattan distance between samples in the training mini-batch within both feature space and target space. As depicted in Fig. 1, it scales the distance between learned features in high-dimensional space by a weighted magnitude of the chronological age difference (see Sec. 3.1). A new ordinality metric is proposed here to quantify the relative ordering of feature representations compared to their actual target label ordering. Results show that our proposed framework preserves ordinality in feature space and improves brain age prediction by a statistically significant amount compared to existing deep learning approaches [23, 9, 21].

One challenge in medical imaging is heterogeneity in the quality of MRI scans due to different scanners and acquisition protocols. Several studies have confined themselves to a single cohort to train and evaluate model performance [23], which could affect multi-site studies or generalization performance. Contrary to that, it is shown that deep learning [34] and machine learning models [16, 15, 24] are not only robust to scanner differences, but diversity in data due to heterogeneous sources can improve model generalization. In this study, we decide to combine cohorts from 5 public data sources to train and validate our model collected from (1) National Alzheimer's Coordinating Center's (NACC), (2) Open Access Series of Imaging Studies (OASIS) [33, 32], (3) International Consortium for Brain Mapping (ICBM), (4) Information eXtraction from Images (IXI), and (5) Autism Brain Imaging Data Exchange-I (ABIDE) [12]. Additionally, disease detection performance is evaluated on an independent dataset. In summary, the main contributions of this research are the following:

1. We formulate Brain Age prediction as an ordinal classification task that outperforms existing regression-based methods by a significant margin.

2. A novel *ORDER loss* is introduced for classification that preserves the ordinality in the learned feature space from target labels, which here is Age.

3. Proposed framework addresses the well-observed issue of systematic bias in predicted biological age from neuroimaging data.

4. Developed model detected subtle differences between clinical groups of Alzheimer's disease, which were not accurately captured by the regression model or other approaches.

2. Related Work

2.1. Neuroimaging based Brain Age prediction

Prior studies on brain age prediction from neuroimaging data [16, 10, 18, 46, 2, 28, 8, 15, 4] use regression techniques such as gaussian process regression, support vector regression, and relevance vector regression. These approaches involve extensive pre-processing of raw structural MRI data and extracting imaging features such as cortical thickness, regional volumes, or surface area using tools such as FreeSurfer or Statistical Parametric Mapping

(SPM). Input to the machine learning models are these pre-processed brain morphological features, and chronological age is the target variable.

More recent studies have also explored deep neural networks to predict brain age using raw neuroimaging data [9, 23, 37, 22, 42, 43, 3] and results demonstrate that deep neural networks outperform traditional machine learning approaches given sufficient training data [3, 9, 21]. Since deep learning methods perform automatic feature extraction from raw structural MRI data, it allows capturing previously unseen imaging signatures related to aging in the brain and makes the model less prone to any biases from pre-processing steps, making it more generalizable.

2.2. Systematic Bias in Predicted Brain Age

In brain age prediction, predicted biological age is often observed to be systematically biased towards the cohort's mean age [27, 26, 44, 45, 5] affected by regression to the mean (RTM) effect, limiting its potential clinical utility. This causes an unexpected overestimation of predicted brain age in young subjects and underestimation among old subjects. Historically, the RTM effect has been attributed to within-subject and between-subject variability [17]. This systematic bias in predicted brain age is not specific to the choice of learning algorithm, data sample imbalance across age groups, or imaging data heterogeneity due to different scanners [27]. Since brain age prediction is traditionally formulated as a regression problem, RTM is a characteristic phenomenon of regression analysis.

Studies that aim to mitigate this systematic bias propose post-hoc correction methods where predicted age is scaled by slope and intercept derived from regression of predicted age or BrainAGE [10, 11, 5] on chronological age. Le *et al*. [26] used chronological age as a covariate when analyzing group-level differences in BrainAGE, whereas Cole *et al*. [10] did not include chronological age in the final adjustment scheme. However, it increased the variance in predicted BrainAGE [5]. Other studies [5, 11] included chronological age in the final age adjustment, but these methods are likely to be inaccurate when the age range of the independent testing dataset differs from the age range of the model's training data. Recently, Zhang *et al*. [47] found that these correction methods do not properly address the systematic bias in predictions. Experiments from that study also show that even though linear [5, 10] and quadratic [44] correction methods push average BrainAGE close to zero, bias in BrainAGE for same-age subjects gets worse.

More fundamentally, correcting the predicted BrainAGE in a two-step process by explicitly controlling for age would make downstream analysis questionable. This highlights the need to develop a direct method that addresses systematic bias in brain age prediction and is more accurate in predicting normal aging.

2.3. Regression as Ordinal Classification

Predicting brain age from imaging data is an ordinal classification task (also known as ordinal regression) since the labels exhibit a natural order. *et al*. [19] conducted a comprehensive exploration of ordinal classification methodologies, categorizing them into three main groups: naive approaches using regression or nominal classification methods, ordinal binary decomposition, and threshold models. However, the efficacy of ordinal decomposition approaches relies heavily on task-specific decomposition strategies, while threshold models demand meticulous calibration of hyperparameters to achieve optimal convergence [39]. In this study, we compare our approach with the nominal classification and regression techniques previously documented in the literature.

In computer vision, it is shown that classification can outperform regression in many tasks, such as age estimation from face images [36, 25, 40], object counting [29], and depth estimation [7]. The target space is discretized into same-size intervals, and surprisingly, models are more accurate in predicting a range of values rather than estimating actual values on a continuous scale. The exact reason for classification outperforming regression has been less explored before. Zhang *et al*. [48] suggest that classification benefits from its ability to learn high entropy feature representations compared to regression, which accounts for the performance gap. Inspired by these insights, we transform the task of brain age prediction from regression to multi-class classification. In brain age prediction, the target output follows a continuous scale consisting of the human life age span. Despite the performance improvement, classification models treat each class label independently from each other, where each wrong prediction is penalized equally. For instance, given a sample with a true age of 53, cross-entropy (CE) penalizes the model by the same magnitude if the wrong prediction was 21 or 52. Hence, the ordinal relationship between target labels is not accurately captured in learned representations of brain age using CE or other loss functions proposed in previous studies [36, 48] (Sec. 4).

One of the initial works that proposed deep learning-based classification for age estimation from facial images was by Rothe *et al*. [40], where they used the expected mean of softmax weights as the estimated age. Pan *et al*. [36] also used softmax expected value for age estimation with an additional mean-variance loss used in training. Mean loss minimizes the difference between the mean of the estimated distribution and the ground truth, while the variance loss minimizes the variance of the estimated distribution, resulting in a concentrated distribution. Different from these approaches, Zhang *et al*. [48] observed that classification allows learning high entropy feature representation with a more diverse feature set compared to regression. They introduce an Euclidean distance-based loss with mean squared error (MSE) loss for regression to increase the marginal entropy such that learned features are spread out while preserving target ordinality. The latter two studies [36, 48] also highlight preserved ordinality in learned feature space from their proposed approaches. However, for brain age prediction, results show that this is not the case when compared to a regression model (Sec. 4).

3. Methods

Fig. 2 gives an overview of our framework for robust brain age prediction. In this section, we first describe our proposed ORDER loss that encodes ordinal information within target labels into learned feature space. Then, in addition to MAE, we define two metrics to measure our model's performance in preserving ordinality and minimizing systematic bias compared to established methods.

3.1. ORDER Loss

To better understand the intuition behind the proposed ORDER loss, we first review the original cross-entropy loss (L_{CE}), which is formulated as:

$$\begin{aligned} L_{CE} &= -\frac{1}{N}\sum_{i=1}^{N}\log(\hat{y}_i) \\ &= -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T x_i}}{\sum_{j=1}^{C} e^{W_{y_j}^T x_i}} \end{aligned} \tag{1}$$

where x_i is input to the last fully connected layer corresponding to i-th sample from training data N, y_i is the hot encoding of the true label, $\hat{y}_i$ is the predicted probability, and $W_{y_j}^T$ is j-th column of last fully connected layer ($j \in [1, C], C$ is number of classes). $W_{y_i}^T x_i$ often denoted as z_i, is the target logit of i-th sample [38].

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \tag{2}$$

In brain age prediction, our main aim is to understand how the dependent variable (age) changes with variations in independent variables (imaging features). Given a sample from class i, cross-entropy loss forces $z_i > z_j (\forall j \neq i)$. However, when the class labels are ordered, it does not guarantee that learned feature representation follows the same order, i.e., $z_i < z_{i+1} < z_{i+2} < ... < z_C$ and $z_i > z_{i-1} > z_{i-2} > ... > z_1$. Even though L_{CE} increases the marginal entropy of feature space, resulting in a diverse feature set, the marginal ordering between class labels is not correctly captured. Keeping the diversity of features from L_{CE} intact, we adjust the target logit z_i with corresponding feature vector x_i's distance to other features $x_j (\forall j \neq i)$ in a batch of samples, weighted by distance between class labels.

$$z' = W_{y_i}^T x_i + \varphi(x_i) \tag{3}$$

where,

$$\varphi(x_i) = \frac{1}{N-1}\sum_{j=1, i\neq j}^{N} |i-j||\bar{x}_i - \bar{x}_j|_{manh} \tag{4}$$

$\bar{x}$ is L_2 normalized vector $\bar{x} = x/max(||x||_2)$. Substituting Eq. 3 in Eq. 2 we get new loss L_T, which can be decomposed into L_{CE} and ORDER loss (L_{ORDER})

$$\begin{aligned} L_T &= -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T x_i + \varphi(x_i)}}{\sum_{j=1}^{C} e^{W_{y_j}^T x_i}} \\ &= -\frac{1}{N}[\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T x_i}}{\sum_{j=1}^{C} e^{W_{y_j}^T x_i}} + \sum_{i=1}^{N}\varphi(x_i)] \\ &= -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T x_i}}{\sum_{j=1}^{C} e^{W_{y_j}^T x_i}} \\ &\quad - \frac{1}{N(N-1)}\sum_{j=1, i\neq j}^{N} |i-j||\bar{x}_i - \bar{x}_j|_{manh} \\ &= L_{CE} + L_{ORDER} \end{aligned} \tag{5}$$

We use the Manhattan distance to calculate the distance between two features x_i and x_j in high-dimensional space. Euclidean distance is the most common metric to measure similarity or distances between two data points. However, Aggarwal *et al.* [1] found that, due to the curse of dimensionality in high-dimensional space, the sparsity of features is significantly high, making them almost equidistant from each other. The ratio between the closest and farthest points from a reference sample approaches 1 in high-dimension space [13]. This further explains the inability of a classification model to capture ordinal information. We explored different orders of distance metrics for ORDER loss, but Manhattan distance performed best (see Sec. 4.5).

3.2. Evaluation Metrics

3.2.1 Measuring Ordinality

To the best of our knowledge, there are no defined metrics in the literature that measure the ordinality of feature representations from a deep learning model with reference to the order of ground truth. Given n images and c ordered classes, we first obtain n features of 512 dimensions from the penultimate layer of a trained model $\{x_1, x_2, ..., x_n\}$. From those features, we calculate c feature centroids $\{f_1, f_2, ..., f_c\}$ using ground-truth labels corresponding to each class. After that, Manhattan distances between f_1 and other feature centroids can be calculated as $D = \{d_{12}, d_{13}, ...d_{1c}\}$ where,

$$d_{ij} = |d_i - d_j|_{manh} \tag{6}$$

Since class labels here are age values in a chronologically increasing order, we get $C = \{1, 2, ..., (c-1)\}$ as the

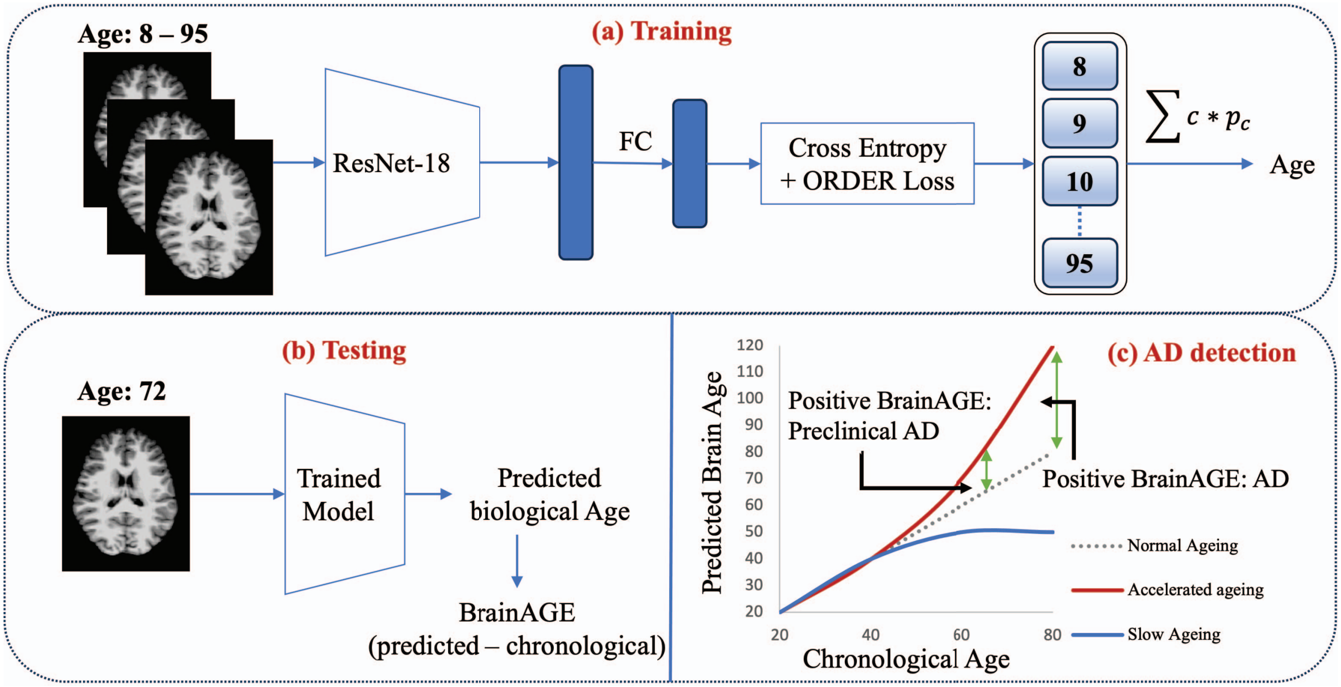


Figure 2. Overview of proposed brain age prediction framework. (a) A 3D ResNet-18 model is trained using lifespan cohort with cross entropy and ORDER losses. Age is calculated as the weighted average of class probabilities from the softmax classifier. (b) At inference, the Brain Age Gap Estimate (BrainAGE) is calculated as the difference between predicted biological age and actual chronological age. (c) The trajectory plot offers a visual interpretation of predicted BrainAGE and its associations with aging patterns. The preclinical AD stage is when the patient behaves cognitively normal, but underlying changes in the brain due to accelerated aging happening at a subtle rate can be captured using BrainAGE.

distance of the first class to others. We define the ordinality metric as the Pearson correlation coefficient between D and C. For a model that perfectly captures ordinal relationships in feature representations, the ordinality score is close to 1. Pearson correlation between two continuous variables measures how much change in one variable is associated with a proportional change in the other variable. Using this metric, our model's performance in capturing age-related order information from labels compared to other approaches is evaluated (see Tab. 3). An ordinality score close to +1 indicates that the learned features have a similar ranking order as their corresponding ground-truth labels and a lower value indicates otherwise.

3.2.2 Quantifying Systematic Bias

Previous approaches discussed in Sec. 3.2.2 that propose post-hoc correction methods use correlation of predicted BrainAGE and chronological age as a measure of underlying systematic bias [26, 27]. Using chronological age to adjust BrainAGE would reduce age dependence on BrainAGE, i.e., $r = 0$. However, it does not address the inherent systematic bias effect caused due to regression. Additionally, this correction method would be questionable when the test dataset does not have the same age range as the training dataset.

To objectively quantify systematic bias caused by regression to the mean effect, we compare the predicted BrainAGE at one standard deviation away from mean [17], i.e., for values less than $(\mu - \sigma)$ and greater than $(\mu + \sigma)$, where μ and σ are mean and standard deviation of target age values of the test set. We refer to these two groups as Systematic Bias - Left and Right (SB-L, SB-R). Since there is an overestimation of predicted biological age in young subjects and an underestimation in old subjects, the bias causes higher BrainAGE and lower BrainAGE values for those respective sub-groups. These scores are compared for different methods, and a value closer to 0 indicates better performance in addressing systematic bias (see Tab. 3).

4. Experiments and Results

We evaluate our proposed brain age prediction framework and other baseline methods on a combined healthy cohort using three different metrics specific to this task. Evaluation metrics include MAE, Ordinality, and Systematic Bias scores.

A 3D ResNet-18 was adopted as the base deep-learning model, and input to the model are 3-dimensional MRI scans with a batch size of 4. Stratified oversampling was employed in classification models, and for regression, samples were stratified based on age groups $(8-12, 12-16, ..., 92-96)$ to perform oversampling. We used AdamW optimizer with a $1e^{-3}$ learning rate and weight decay of $1e^{-2}$. Each model was trained for 100 epochs and with early stopping to avoid over-fitting. All experiments were performed on NVIDIA's $A100$ $80GB$ GPUs to train, validate, and test the models.

4.1. Datasets and Preprocessing

Since most medical imaging datasets are part of multi-center studies, differences in scanners, imaging protocols, variations in vendors, and their hardware account for heterogeneity in data. Deep learning models are known to be robust against heterogeneity in data. In fact, including more heterogeneous data in model training improves its generalization on out-of-distribution data [34]. With that consideration, a combined lifespan cohort of $7,377$ T1-weighted MRI scans of healthy participants collected from five different public sources was used in model training.

Lifespan cohort: All the age prediction models were trained, validated, and tested on a healthy cohort (age: 8-95 years) collected from (1) NACC Uniform Data Set (UDS) from 1999 to March 2021 (2) OASIS (3) ICBM (4) IXI (http://brain-development.org/ixidataset/) and (5) ABIDE. These cohorts included both 1.5T and 3T scans with predominantly Caucasian participants but also included other race/ethnic groups. The number of samples and age range per cohort are summarized in Tab. 1.

All five cohorts were preprocessed using an in-house data preprocessing pipeline. T1-weighted MR images were first aligned to the MNI template with rigid transformation, and then intensity normalized and conformed using FreeSurfer v7 to generate preprocessed images at 1 mm isotropic voxels with a 256 x 256 x 256 matrix.

Discovery cohort: Additionally, $1,584$ MRI scans were collected from the Alzheimer's Disease Neuroimaging Initiative (ADNI; https://adni.loni.usc.edu/) database containing a mix of healthy, cognitively impaired, and AD patients. This cohort was used as an independent testing and discovery dataset to evaluate model performance in predicting age and its ability to differentiate clinical groups in AD. Priority was given to scans with matching PET data and participants who had longitudinal follow-ups. For healthy controls (HCs), a random subset was selected from the overall ADNI set and included in this analysis. The diagnostic status was determined based on ADNI clinical data. In this analysis, HC (N=678) participants had normal cognition and did not convert to MCI or AD in follow-up visits. HC to MCI converters (HC-MCI, N=179) are participants who had normal cognition at baseline but converted to MCI during follow-up. MCI-stable (MCIs, N=432) participants had a baseline diagnosis of MCI and stayed unchanged in follow-ups. MCI to AD converters (MCI-AD, N=139) are those participants with MCI diagnosis at baseline and subsequently converted to AD. AD (N=156) patients are those who were diagnosed with AD at baseline.

Dataset	**Count**	**Age Range** (yrs)	**Mean** ± **STD**
NACC	4,132	18 - 95	67.5 ± 10.8
OASIS	1,432	8 - 94	27.9 ± 20.7
ICBM	1,101	18 - 80	37.6 ± 15.4
IXI	536	20 - 86	48.8 ± 16.5
ABIDE	176	18 - 56	26.1 ± 7.0
ADNI	1,584	55 - 98	73.3 ± 7.3

Table 1. Age range with distribution and number of samples for each cohort. The lifespan cohort comprises NACC, OASIS, ICBM, IXI, and ABIDE, whereas the Discovery cohort consists of samples from the ADNI cohort.

4.2. BrainAGE prediction

We compare our proposed method's performance in predicting the brain age of healthy individuals from lifespan cohort to four baseline methods, including two regression and two classification models (Tab. 2). For classification models, age values were rounded off to the closest integer and assigned respective class labels. Only $535(7.3\%)$ samples from the lifespan cohort had non-integer age values.

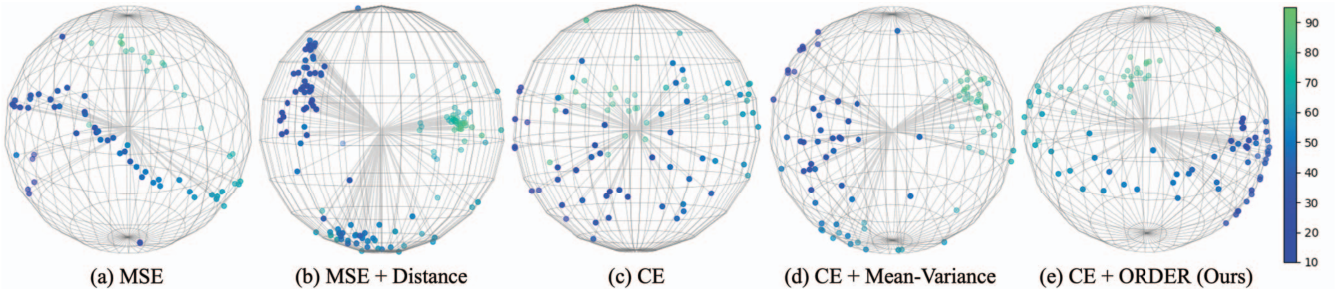
Our model performed best on the healthy test set with MAE 2.56, outperforming standard MSE and cross-entropy loss models. Among other competing methods, the classification model with mean-variance loss performed best. The model with cross-entropy loss outperforms the MSE model due to its ability to learn high entropy features (Fig. 3), where inter-class features are spread out, and intra-class features are compact [6]. Surprisingly, adding Euclidean distance-based regularizer to MSE loss did not improve the regression model's performance. Our method's performance is also significantly better than MAE reported by prior studies using regression analysis [21, 23, 9], however, on different cohorts.

Method	**MAE**
MSE	3.93
MSE + Distance [48]	4.57
CE [40]	3.33
CE + Mean-Variance [36]	2.65
CE + ORDER (Ours)	**2.56**

Table 2. Brain age prediction results on lifespan cohort. MAE measures the difference between predicted and actual chronological age on the same test set. **Bold** numbers represent the best results, while underlined numbers represent second-best results.

4.3. Ordinality and Systematic Bias

We further evaluate our model's ability to preserve ordinality and address systematic bias in predicted BrainAGE using metrics defined in Sec. 3.2.1 and Sec. 3.2.2. As ex-



(a) MSE (b) MSE + Distance (c) CE (d) CE + Mean-Variance (e) CE + ORDER (Ours)

Figure 3. t-SNE visualization of embeddings from models' penultimate layer: (a) When using MSE loss, embeddings maintain ordinal relationships but are tightly packed, resulting in a low-entropy feature space (b) MSE with Euclidean distance loss spreads out embeddings but struggles to preserve ordinal relationships accurately (c) Cross-entropy (CE) further spreads embeddings, creating a high-entropy space, but at the cost of losing ordinal information (d) Mean-variance loss combined with cross-entropy creates a high-entropy feature space and slightly improves ordinality (Tab. 3). (e) ORDER loss combined with cross-entropy achieves the best balance: it accurately preserves ordinality, maintains a high-entropy space, and improves overall performance. Embeddings are colored-coded based on their ground truth age values $[10-95]$.

pected, the model with MSE loss had the highest ordinality score (Tab. 3). Our classification model with ORDER loss had an ordinality score much closer to standard MSE loss than other methods, demonstrating its effectiveness in learning ordinal information. Fig. 3 offers a visual comparison of learned feature space using different loss functions to confirm this further.

Furthermore, the model with ORDER loss also performed best in reducing systematic bias measured by average BrainAGE values at one standard deviation away from the mean. The mean of the test set was 53.4 with a standard deviation of 22.2. Hence, the bias scores reported in Tab. 3 are BrainAGE values for age < 31.2 (SB-L) and age > 75.6 (SB-R). Values closer to zero reflect a better reduction in systematic bias. Both MSE-based models had a higher systematic bias due to the inherent RTM effect. Due to its ability to learn class-specific and diverse feature sets, cross-entropy loss reduces bias effects for SB-L and SB-R groups. Incorporating order information allows the model to learn the relative ranking of labels, further improving ordinal classification performance.

Method	Ordinality	Systematic Bias	
		SB-L	SB-R
MSE	**0.99**	3.4	-4.2
MSE + Distance	0.95	4.8	-4.1
CE	0.31	1.1	-3.6
CE + Mean-Variance	0.58	0.4	-4.2
CE + ORDER	0.98	**0.1**	**-2.5**

Table 3. Performance evaluation of all methods in preserving ordinality and addressing systematic bias in brain age prediction using metrics defined in Sec. 3.2.1 and Sec. 3.2.2.

4.4. Alzheimer's Disease detection

AD has a prolonged preclinical phase where brain changes manifest subtly as accelerated aging [30]. Fig. 2 illustrates this phase, showing accelerated aging diverging slightly from normal aging. MCI, a pre-dementia stage, involves greater cognitive decline than typical aging [41]. BrainAGE can help detect and monitor this stage early.

The discovery cohort (Sec. 4.1) obtained from ADNI with five clinical groups was used to test BrainAGE prediction using different methods. Trained models were applied to this cohort using the abovementioned methods to calculate BrainAGE. These five groups were ranked $[1-5]$ in an increasing order of disease severity as HC $<$ HC-MCI $<$ MCI-stable $<$ MCI-AD $<$ AD. Since disease severity is proportional to accelerated aging, we expect the average predicted BrainAGE to follow the same order. Pearson correlation is calculated between the model's predicted BrainAGE and rank of disease severity. A high correlation would indicate the model's ability to accurately characterize aging signatures along the AD continuum via estimated BrainAGE. From Tab. 4, we see that only the model with MSE and our proposed loss have a high correlation.

We further compare the ability of MSE and ORDER loss models to detect subtle differences between these clinical groups accurately. Fig. 4 shows the MSE model had a more disruptive trend in predicted BrainAGEs between groups, i.e., there was a higher difference between AD and MCI-AD $(p = 0.16)$ compared to AD and MCI-stable $(p = 0.56)$. Whereas the ORDER loss model had an overall consistent trend in statistical significance between groups associated with actual disease severity, highlighting its better discriminative power. It was also able to better detect differences between HC and HC-MCI subjects $(p = 0.07)$ compared to MSE $(p = 0.34)$, which is crucial for early AD detection. Although our model's performance wasn't as strong as MSE in distinguishing between HC-MCI and MCI-stable, we posit that this could be attributed to the definitions of clinical groups used here. The absence of clinical tools to definitively differentiate HC-MCI from MCI-stable groups, given that subjects exhibit normal cognitive behavior and no discernible symptoms despite age-related brain changes, might contribute to this outcome. We plan to work with clinicians to further investigate these observations from both groups.

Method	HC	HC-MCI	MCIs	MCI-AD	AD	Corr.
MSE	-1.2	-0.8	-0.3	0.8	1.5	0.98
MSE + Distance	-2.7	-1.9	-1.7	-0.9	0.9	0.94
CE	-1.9	-1.5	-3.4	-2.3	-4.1	-0.75
CE + MV	-1.6	-0.3	-0.5	0.8	2.8	0.94
CE + ORDER	-1.5	-0.8	-0.3	1.2	2.0	0.98

Table 4. Average BrainAGE values across the five clinical groups of AD. Last column is the Pearson correlation between average BrainAGE values and disease severity of clinical groups in increasing order from HC to AD. MV: Mean-Variance.

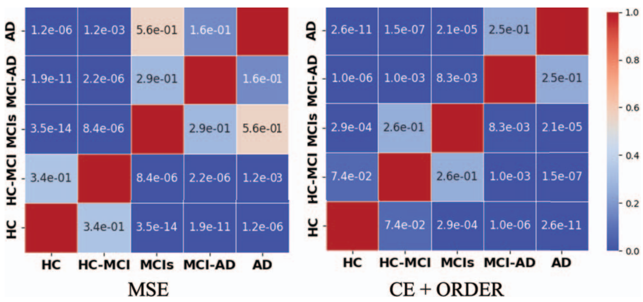


Figure 4. Heatmap of statistical significances between the five clinical groups of AD calculated as p values from a t-test on predicted BrainAGE from respective groups, for MSE and cross-entropy with ORDER loss models.

4.5. Ablation studies

Distance Metric: We explored different L_k norm distance metrics for ORDER loss and found Manhattan distance best performing across all evaluation measures. L_k norm distance between two points x and y in high-dimensional space, given $(x, y \in R^d)$, is be defined as:

$$L_k(x, y) = \sum_{i=1}^{d} [||x^i - y^i||^k]^{\frac{1}{k}} \tag{7}$$

Aggarwal *et al.* [1] showed that Manhattan distance $(k = 1)$ is a more suitable distance metric than Euclidean $(k = 2)$ for high-dimensional data. They recommended using $k \leq 1$ to improve downstream classification performance. Later studies showed that fractional distance metrics, i.e., $(k < 1)$, do not systematically address the issue of the curse of dimensionality [35] but should be a choice depending on the training data distribution. For the high-dimensional neuroimaging dataset used here, results indicate that Manhattan distance is more accurate in preserving ordinality and improving class separability compared to Euclidean or fractional distance metrics (see Tab. 5).

ORDER loss with Classification vs. Regression: We experimented with the proposed ORDER loss using both classification and regression frameworks. As discussed in the paragraph above, since Euclidean and Manhattan distances performed significantly better than fractional distances, we explored regression models with our loss for $k = \{1, 2\}$ (Tab. 5). Results show that distance-based regularization does not work well in regression models. Our model with cross-entropy loss and Manhattan distance-based ordinal regularization performed best across the three metrics.

k	Loss	MAE	Ordinality	Systematic Bias	
				SB-L	SB-R
1/2	CE	6.05	0.85	5.31	-5.19
2/3	CE	18.51	0.13	30.67	-28.27
1	**CE**	**2.56**	**0.98**	**0.11**	**-2.5**
1	MSE	4.66	0.95	2.19	-4.98
2	CE	2.90	0.10	0.93	-3.04
2	MSE	4.57	0.95	4.83	-4.13

Table 5. Ablation studies on the proposed framework components evaluated by MAE, ordinality, and systematic bias scores. k denotes different L_k-norm distance metrics defined in Eq. 7

.

5. Conclusion

This paper proposes a novel ordinal-distance regularization loss for robust brain age prediction using deep learning. ORDER loss in an ordinal classification framework outperforms regression-based brain age prediction methods, reduces systematic bias in predictions, and preserves ordinality in learned feature space. Improved performance is attributed to ordering information encoded in the model using ORDER loss and the ability of cross entropy loss to learn high entropy feature representations. The predicted BrainAGE from this model is a more reliable imaging biomarker for diagnosing AD and predicting its early onset. We believe this framework can be generalized to other regression tasks to improve prediction and address the RTM effect if present, which we aim to investigate further in future work.

Acknowledgments. This research received support from the National Institute on Aging of the National Institutes of Health under Award Numbers R01AG069453, P30AG072980, RF1AG073424, Banner Alzheimer's Foundation and the Arizona Department of Health Services to Arizona Alzheimer's Research Center. We thank Arizona State University Research Computing (ASURC) for hosting and maintaining our computing resources.

References

[1] Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. On the surprising behavior of distance metrics in high dimensional space. In *Database Theory—ICDT 2001: 8th International Conference London, UK, January 4–6, 2001 Proceedings 8*, pages 420–434. Springer, 2001.

[2] Lea Baecker, Jessica Dafflon, Pedro F Da Costa, Rafael Garcia-Dias, Sandra Vieira, Cristina Scarpazza, Vince D Calhoun, Joao R Sato, Andrea Mechelli, and Walter HL Pinaya. Brain age prediction: A comparison between machine learning models using region-and voxel-based morphometric data. *Human brain mapping*, 42(8):2332–2346, 2021.

[3] Vishnu M Bashyam, Guray Erus, Jimit Doshi, Mohamad Habes, Ilya M Nasrallah, Monica Truelove-Hill, Dhivya Srinivasan, Liz Mamourian, Raymond Pomponio, Yong Fan, et al. Mri signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide. *Brain*, 143(7):2312–2324, 2020.

[4] Iman Beheshti, Shiwangi Mishra, Daichi Sone, Pritee Khanna, and Hiroshi Matsuda. T1-weighted mri-driven brain age estimation in alzheimer's disease and parkinson's disease. *Aging and disease*, 11(3):618, 2020.

[5] Iman Beheshti, Scott Nugent, Olivier Potvin, and Simon Duchesne. Bias-adjustment in neuroimaging-based brain age frameworks: A robust scheme. *NeuroImage: Clinical*, 24:102063, 2019.

[6] Malik Boudiaf, Jérôme Rony, Imtiaz Masud Ziko, Eric Granger, Marco Pedersoli, Pablo Piantanida, and Ismail Ben Ayed. A unifying mutual information view of metric learning: cross-entropy vs. pairwise losses. In *European conference on computer vision*, pages 548–564. Springer, 2020.

[7] Yuanzhouhan Cao, Zifeng Wu, and Chunhua Shen. Estimating depth from monocular images as classification using deep fully convolutional residual networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(11):3174–3182, 2017.

[8] James H Cole, Robert Leech, David J Sharp, and Alzheimer's Disease Neuroimaging Initiative. Prediction of brain age suggests accelerated atrophy after traumatic brain injury. *Annals of neurology*, 77(4):571–581, 2015.

[9] James H Cole, Rudra PK Poudel, Dimosthenis Tsagkrasoulis, Matthan WA Caan, Claire Steves, Tim D Spector, and Giovanni Montana. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage*, 163:115–124, 2017.

[10] James H Cole, Stuart J Ritchie, Mark E Bastin, Valdés Hernández, S Muñoz Maniega, Natalie Royle, Janie Corley, Alison Pattie, Sarah E Harris, Qian Zhang, et al. Brain age predicts mortality. *Molecular psychiatry*, 23(5):1385–1392, 2018.

[11] Ann-Marie G de Lange, Tobias Kaufmann, Dennis van der Meer, Luigi A Maglanoc, Dag Alnæs, Torgeir Moberget, Gwenaëlle Douaud, Ole A Andreassen, and Lars T Westlye. Population-based neuroimaging reveals traces of childbirth in the maternal brain. *Proceedings of the National Academy of Sciences*, 116(44):22341–22346, 2019.

[12] Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, Mirella Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659–667, 2014.

[13] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

[14] Kristy Draper and Jennie Ponsford. Cognitive functioning ten years following traumatic brain injury and rehabilitation. *Neuropsychology*, 22(5):618, 2008.

[15] Katja Franke and Christian Gaser. Longitudinal changes in individual brainage in healthy aging, mild cognitive impairment, and alzheimer's disease. *GeroPsych*, 2012.

[16] Katja Franke, Gabriel Ziegler, Stefan Klöppel, Christian Gaser, Alzheimer's Disease Neuroimaging Initiative, et al. Estimating the age of healthy subjects from t1-weighted mri scans using kernel methods: exploring the influence of various parameters. *Neuroimage*, 50(3):883–892, 2010.

[17] MJ Gardner and JA Heady. Some effects of within-person variability in epidemiological studies. *Journal of Chronic Diseases*, 26(12):781–795, 1973.

[18] Christian Gaser, Katja Franke, Stefan Klöppel, Nikolaos Koutsouleris, Heinrich Sauer, and Alzheimer's Disease Neuroimaging Initiative. Brainage in mild cognitive impaired patients: predicting the conversion to alzheimer's disease. *PloS one*, 8(6):e67346, 2013.

[19] Pedro Antonio Gutiérrez, Maria Perez-Ortiz, Javier Sanchez-Monedero, Francisco Fernandez-Navarro, and Cesar Hervas-Martinez. Ordinal regression methods: survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):127–146, 2015.

[20] Mary N Haan and Robert Wallace. Can dementia be prevented? brain aging in a population-based context. *Annu. Rev. Public Health*, 25:1–24, 2004.

[21] Koichi Ito, Ryuichi Fujimoto, Tzu-Wei Huang, Hwann-Tzong Chen, Kai Wu, Kazunori Sato, Yasuyuki Taki, Hiroshi Fukuda, and Takafumi Aoki. Performance evaluation of age estimation from t1-weighted images using brain local features and cnn. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 694–697. IEEE, 2018.

[22] Huiting Jiang, Na Lu, Kewei Chen, Li Yao, Ke Li, Jiacai Zhang, and Xiaojuan Guo. Predicting brain age of healthy adults based on structural mri parcellation using convolutional neural networks. *Frontiers in neurology*, 10:1346, 2020.

[23] Benedikt Atli Jónsson, Gyda Bjornsdottir, TE Thorgeirsson, Lotta María Ellingsen, G Bragi Walters, DF Gudbjartsson, Hreinn Stefansson, Kari Stefansson, and MO Ulfarsson. Brain age prediction using deep learning uncovers associated sequence variants. *Nature communications*, 10(1):5409, 2019.

[24] Tobias Kaufmann, Dennis van der Meer, Nhat Trung Doan, Emanuel Schwarz, Martina J Lund, Ingrid Agartz, Dag Alnæs, Deanna M Barch, Ramona Baur-Streubel, Alessandro

Bertolino, et al. Common brain disorders are associated with heritable patterns of apparent aging of the brain. *Nature neuroscience*, 22(10):1617–1623, 2019.

[25] Andreas Lanitis, Chrisina Draganova, and Chris Christodoulou. Comparing different classifiers for automatic age estimation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1):621–628, 2004.

[26] Trang T Le, Rayus T Kuplicki, Brett A McKinney, Hung-Wen Yeh, Wesley K Thompson, Martin P Paulus, and Tulsa 1000 Investigators. A nonlinear simulation framework supports adjusting for age when analyzing brainage. *Frontiers in aging neuroscience*, 10:317, 2018.

[27] Hualou Liang, Fengqing Zhang, and Xin Niu. Investigating systematic bias in brain age estimation with application to post-traumatic stress disorders. Technical report, Wiley Online Library, 2019.

[28] Franziskus Liem, Gaël Varoquaux, Jana Kynast, Frauke Beyer, Shahrzad Kharabian Masouleh, Julia M Huntenburg, Leonie Lampe, Mehdi Rahim, Alexandre Abraham, R Cameron Craddock, et al. Predicting brain-age from multimodal imaging data captures cognitive impairment. *Neuroimage*, 148:179–188, 2017.

[29] Liang Liu, Hao Lu, Haipeng Xiong, Ke Xian, Zhiguo Cao, and Chunhua Shen. Counting objects by blockwise classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10):3513–3527, 2019.

[30] Justin M Long and David M Holtzman. Alzheimer disease: an update on pathobiology and treatment strategies. *Cell*, 179(2):312–339, 2019.

[31] Carlos López-Otín, Maria A Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. The hallmarks of aging. *Cell*, 153(6):1194–1217, 2013.

[32] Daniel S Marcus, Anthony F Fotenos, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies: longitudinal mri data in nondemented and demented older adults. *Journal of cognitive neuroscience*, 22(12):2677–2684, 2010.

[33] Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007.

[34] Gustav Mårtensson, Daniel Ferreira, Tobias Granberg, Lena Cavallin, Ketil Oppedal, Alessandro Padovani, Irena Rektorova, Laura Bonanni, Matteo Pardini, Milica G Kramberger, et al. The reliability of a deep learning model in clinical out-of-distribution mri data: a multicohort study. *Medical Image Analysis*, 66:101714, 2020.

[35] Evgeny M Mirkes, Jeza Allohibi, and Alexander Gorban. Fractional norms and quasinorms do not help to overcome the curse of dimensionality. *Entropy*, 22(10):1105, 2020.

[36] Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen. Mean-variance loss for deep age estimation from a face. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5285–5294, 2018.

[37] Han Peng, Weikang Gong, Christian F Beckmann, Andrea Vedaldi, and Stephen M Smith. Accurate brain age prediction with lightweight deep neural networks. *Medical image analysis*, 68:101871, 2021.

[38] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

[39] Riccardo Rosati, Luca Romeo, Víctor Manuel Vargas, Pedro Antonio Gutiérrez, César Hervás-Martínez, and Emanuele Frontoni. A novel deep ordinal classification approach for aesthetic quality control classification. *Neural Computing and Applications*, 34(14):11625–11639, 2022.

[40] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2-4):144–157, 2018.

[41] Dennis J Selkoe. Alzheimer's disease–genotypes, phenotype, and treatments. *Science*, 275(5300):630–631, 1997.

[42] Jay Shah, Valentina Ghisays, Yinghua Chen, Ji Luo, Baoxin Li, Eric M Reiman, Kewei Chen, Teresa Wu, and Yi Su. Mri signatures of brain age in the alzheimer's disease continuum. *Alzheimer's & Dementia*, 18:e061942, 2022.

[43] Jay Shah, Ji Luo, Javad Sohankar, Eric M Reiman, Kewei Chen, Yi Su, Baoxin Li, and Teresa Wu. A multi-class deep learning model to estimate brain age while addressing systematic bias of regression to the mean. In *Alzheimer's Association International Conference*. ALZ, 2023.

[44] Stephen M Smith, Diego Vidaurre, Fidel Alfaro-Almagro, Thomas E Nichols, and Karla L Miller. Estimation of brain age delta from brain imaging. *Neuroimage*, 200:528–539, 2019.

[45] Matthias S Treder, Jonathan P Shock, Dan J Stein, StéFan Du Plessis, Soraya Seedat, and Kamen A Tsvetanov. Correlation constraints for regression models: Controlling bias in brain age prediction. *Frontiers in psychiatry*, 12:615754, 2021.

[46] SA Valizadeh, Jürgen Hänggi, Susan Mérillat, and Lutz Jäncke. Age prediction on the basis of brain anatomical measures. *Human brain mapping*, 38(2):997–1008, 2017.

[47] Biao Zhang, Shuqin Zhang, Jianfeng Feng, and Shihua Zhang. Age-level bias correction in brain age prediction. *NeuroImage: Clinical*, 37:103319, 2023.

[48] Shihao Zhang, Linlin Yang, Michael Bi Mi, Xiaoxu Zheng, and Angela Yao. Improving deep regression with ordinal entropy. *arXiv preprint arXiv:2301.08915*, 2023.