frontiers in Neuroscience

ORIGINAL RESEARCH
published: 28 May 2021
doi: 10.3389/fnins.2021.674055

Explainable Deep Learning for Personalized Age Prediction With Brain Morphology

Angela Lombardi[1,2], *Domenico Diacono*[2]*, *Nicola Amoroso*[2,3], *Alfonso Monaco*[2], *João Manuel R. S. Tavares*[4], *Roberto Bellotti*[1,2†] *and Sabina Tangaro*[2,5†]

[1] *Dipartimento di Fisica, Universitá degli Studi di Bari Aldo Moro, Bari, Italy,* [2] *Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Bari, Italy,* [3] *Dipartimento di Farmacia - Scienze del Farmaco, Università degli Studi di Bari Aldo Moro, Bari, Italy,* [4] *Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial, Departamento de Engenharia Mecânica, Faculdade de Engenharia, Universidade Do Porto, Porto, Portugal,* [5] *Dipartimento di Scienze del Suolo, della Pianta e degli Alimenti, Università degli Studi di Bari Aldo Moro, Bari, Italy*

OPEN ACCESS

Edited by:
John Ashburner,
University College London,
United Kingdom

Reviewed by:
Islem Rekik,
Istanbul Technical University, Turkey
James H. Cole,
University College London,
United Kingdom
Gidon Levakov,
Ben-Gurion University of the Negev,
Israel

***Correspondence:**
Domenico Diacono
domenico.diacono@ba.infn.it

† *These authors share last authorship*

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *10 March 2021*
Accepted: *26 April 2021*
Published: *28 May 2021*

Citation:
Lombardi A, Diacono D, Amoroso N, Monaco A, Tavares JMRS, Bellotti R and Tangaro S (2021) Explainable Deep Learning for Personalized Age Prediction With Brain Morphology. Front. Neurosci. 15:674055. doi: 10.3389/fnins.2021.674055

Predicting brain age has become one of the most attractive challenges in computational neuroscience due to the role of the predicted age as an effective biomarker for different brain diseases and conditions. A great variety of machine learning (ML) approaches and deep learning (DL) techniques have been proposed to predict age from brain magnetic resonance imaging scans. If on one hand, DL models could improve performance and reduce model bias compared to other less complex ML methods, on the other hand, they are typically black boxes as do not provide an in-depth understanding of the underlying mechanisms. Explainable Artificial Intelligence (XAI) methods have been recently introduced to provide interpretable decisions of ML and DL algorithms both at local and global level. In this work, we present an explainable DL framework to predict the age of a healthy cohort of subjects from ABIDE I database by using the morphological features extracted from their MRI scans. We embed the two local XAI methods SHAP and LIME to explain the outcomes of the DL models, determine the contribution of each brain morphological descriptor to the final predicted age of each subject and investigate the reliability of the two methods. Our findings indicate that the SHAP method can provide more reliable explanations for the morphological aging mechanisms and be exploited to identify personalized age-related imaging biomarker.

Keywords: explainable artificial intelligence, XAI, brain aging, deep neural networks, machine learning, MRI, FreeSurfer, morphological features

1. INTRODUCTION

Brain age prediction has become a challenging topic in computational neuroscience, due to the strong link between aging processes and several brain disorders and diseases (Franke and Gaser, 2012; Gaser et al., 2013; Koutsouleris et al., 2014; Cole and Franke, 2017b; Wang et al., 2019). Accordingly, accurate age prediction models measuring the difference between the chronological age and the predicted brain age (i.e., the age gap) have been developed to help identifying novel functional and structural biomarkers for such diseases and provide systems for early diagnosis (Cole et al., 2015, 2019; Cole and Franke, 2017a). In particular, machine learning (ML) and deep learning (DL) algorithms have been successfully applied to predict age from brain MRI scans. Two

main approaches are largely adopted to perform brain age prediction: on one hand, a number of selected features such as morphological descriptors, graph-based or other imaging-related features can be extracted from imaging to train different models (Erus et al., 2015; Amoroso et al., 2018, 2019; Bellantuono et al., 2020; Han et al., 2020); on the other hand, more complex models such as convolutional neural networks directly exploiting raw image as input have proven to be particularly effective in brain age prediction even in broad age ranges (Cole et al., 2017, 2019; Feng et al., 2020; Levakov et al., 2020; Peng et al., 2021). Although convolutional neural networks offer undoubted advantages such as reduced preprocessing time and high performance (Cole et al., 2017), both ML and DL feature-based learning approaches based on morphological features are still widely adopted by scientific communities as they allow to investigate the morphological age-related brain changes in a great variety of disorders and conditions (Van Rooij et al., 2018; Corps and Rekik, 2019; Boedhoe et al., 2020; Han et al., 2020).

Several works have shown that DL models improve performance and reduce model bias compared to other less complex ML methods (Couvy-Duchesne et al., 2020; Da Costa et al., 2020; Lombardi et al., 2020c); however, current DL approaches applied to neuroimaging typically do not provide an in-depth understanding of the underlying mechanisms and how they contributed to the outcome. Understanding how the models affect the decisions and how each feature is related to the outcomes can increase confidence in the models and broaden their applications in the clinical setting (Carvalho et al., 2019; Holzinger et al., 2019). In order to overcome these limitations, new explainable methods have been introduced in the last 5 years. Explainable Artificial Intelligence (XAI) is a relatively new field of Artificial Intelligence and it comprises a large amount of techniques that combines ML algorithms with explanatory techniques to develop explainable solutions that have been extensively applied in different domains (Gunning, 2017; Adadi and Berrada, 2018; Biecek, 2018; Guidotti et al., 2018; Miller, 2019; Arrieta et al., 2020; Bussmann et al., 2020). Recent work has suggested that XAI methods constitute a fundamental pillar for personalized medicine, including individualized interventions and targeted treatments (Vu et al., 2018; Fellous et al., 2019; Langlotz et al., 2019). Most widespread explainable techniques comprise local model-agnostic methods that focus on explaining individual predictions of any ML models, such as LIME (Ribeiro et al., 2016, 2018) and SHAP (Lundberg and Lee, 2017). These methods aim at estimating the contribution of individual features toward a specific prediction by perturbing a given instance and observing the effect of these perturbations on the output of the model.

However, as far as we know, there has been little analysis of the reliability and robustness of the explanation methods in computational neuroscience, making their utility for critical applications unclear. In this work, we present an explainable DL framework to predict the age of a healthy cohort of subjects from ABIDE I database (Di Martino et al., 2014) by using morphological features extracted from their MRI scans. We embed two local XAI methods to explain the outcomes of the DL models and determine the contribution of each brain morphological descriptor to the final predicted age of each subject. We propose a complete architecture to compare the two methods, determine their reliability and to extract information on the importance of the most age-related morphological descriptors in order to encourage the use of DL models in clinical settings.

2. MATERIALS

2.1. Subjects

In this study, we exploited the same dataset used in our previous work (Lombardi et al., 2020b). In particular, we selected $T = 378$ T1-weighted MRI publicly available scans of a cohort of typically-developing individuals from the Autism Brain Imaging Data Exchange (ABIDE I) collected from 17 international sites. The T1-weighted MRI scans were collected with 3 Tesla scanners with different characteristics such as manufacturers and parameters (e.g., echo time, repetition time, flip angle, and field of view). More details about images and acquisition protocols from each site are available at the web page of the initiative[1]. All participating sites received local Institutional Review Board approval for acquisition of the contributed data. Only male subjected were considered in our analysis due to the high imbalance between male and female subjects in the ABIDE data sample. Additionally, we used the full IQ (FIQ) test scores from the phenotype information file and the Signal to Noise Ratio (SNR) from the anatomical quality assessment metrics provided by the publicly available ABIDE Preprocessed repository (Craddock et al., 2013). The SNR was computed as the mean intensity within gray matter divided by the standard deviation of the values outside the brain (Magnotta et al., 2006). The demographic and imaging-related characteristics of the studied subjects are listed in **Table 1** for each of the 17 sites.

2.2. Morphological Features

As in our previous work (Lombardi et al., 2020b), the T1 raw scans were preprocessed by using the recon-all pipeline from the software FreeSurfer v.5.3.0 (Dale et al., 1999; Fischl et al., 1999, 2002) on ReCaS datacenter[2] (Lombardi et al., 2019). The recon-all pipeline allows to segment the brain into 68 cortical regions and 40 sub-cortical region by means of the Desikan–Killiany atlas (Desikan et al., 2006) and Aseg Atlas (Fischl et al., 2002). The output of the pipeline consists in several statistical morphological features related to surface, curvature, thickness and white matter volumes of the cortical regions and volumes of the sub-cortical regions as well as some global brain metrics including surface and volume statistics of each hemisphere, total cerebellar gray and white matter volume, brainstem volume, corpus callosum volume, white matter hypointensities. More details about the steps performed can be found at the web page of the pipeline[3] and in our previous work (Lombardi et al., 2020b). We constructed the matrix of the features of dimension $T \times P$ with $T = 387$, and

[1] http://fcon_1000.projects.nitrc.org/indi/abide
[2] https://www.recas-bari.it/index.php/en/
[3] https://surfer.nmr.mgh.harvard.edu/fswiki/recon-all

$P = 1,213$, where each row represents a single subject described by P morphological features.

3. METHODS

In this study, we developed a DL framework to:

- Predict the brain age of a healthy cohort of subjects by using their morphological features and DNN models;
- Exploit two local XAI methods to extract personalized age-related features;
- Investigate the reliability of these individual age-related features;
- Compare the two XAI methods.

TABLE 1 | Demographic and imaging-related information of the subjects per site.

Site	Samples	Age range (years)	FIQ (mean ± std)	SNR (mean ± std)
CMU	2	21 – 25	109.5 ± 0.7	42.5 ± 6.4
KKI	23	8 – 13	112.9 ± 9.4	23.9 ± 7.8
Leuven 1	13	18 – 29	116.5 ± 12.7	16 ± 1.7
Leuven 2	14	12 – 17	NA	13.5 ± 1.7
MaxMun	24	7 – 48	112.1 ± 8.8	20.2 ± 3.5
NYU	77	6 – 32	113.7 ± 12.6	12.6 ± 1.6
Olin	13	10 – 23	116.3 ± 17.0	18.3 ± 2.5
Pitt	22	12 – 33	110.4 ± 8.3	9.4 ± 1.6
SBL	14	20 – 42	NA	6.3 ± 1.2
SDSU	14	12 – 17	110.5 ± 10.3	20.5 ± 4.6
Trinity	25	12 – 25	110.8 ± 12.2	11.3 ± 2.9
UCLA 1	28	9 – 18	104.6 ± 10.6	13.8 ± 1.9
UCLA 2	11	10 – 14	113.1 ± 11.4	13.8 ± 2.1
UM 1	32	8 – 19	109.8 ± 8.7	22.7 ± 6.5
UM 2	17	13 – 29	110.3 ± 10.2	24.3 ± 5.0
USM	43	8 – 40	115.1 ± 13.7	20.5 ± 2.0
Yale	6	8 – 17	108.1 ± 13.3	21.5 ± 10.8
Total	378	6 – 48	112.1 ± 11.7	16.6 ± 6.6

The overall proposed framework is shown in **Figure 1**. We adopted a leave-one-site cross validation regression scheme: the data from one site are used as a test set to evaluate the performance of the model while the data from all the other sites are used as training set. This cross-validation scheme has been extensively used in multisite studies as it is possible to test the generalization of the models to a new site, and to investigate the correlation between the variability of the characteristics of the different sites and the performance of the models (Abraham et al., 2017; Bhaumik et al., 2018; Heinsfeld et al., 2018). Since in general the ML algorithms can be sensitive with respect to changes in the training set, returning both the performance and the feature ranking varying from round to round, for each cross-validation round, we randomly under-sampled the training set $N = 100$ times by selecting the 80% of the samples to produce small variations of the composition of the set and for each iteration we trained a DNN model to predict the chronological age of the subjects Y, using a fixed percentage of the samples to perform the tuning of the parameters. We tested the DNN models on each sample of the test set collecting $N = 100$ performance MAE values for each subject. Moreover, we applied both SHAP and LIME algorithms to extract the age-related feature importance vector for each subject collecting the two matrix S and L of dimension $[N \times P]$, whose generic element s_{nk} (l_{nk}) indicates the SHAP (LIME) value for the k feature within the n iteration. Accordingly, we analyzed the resulted matrices to

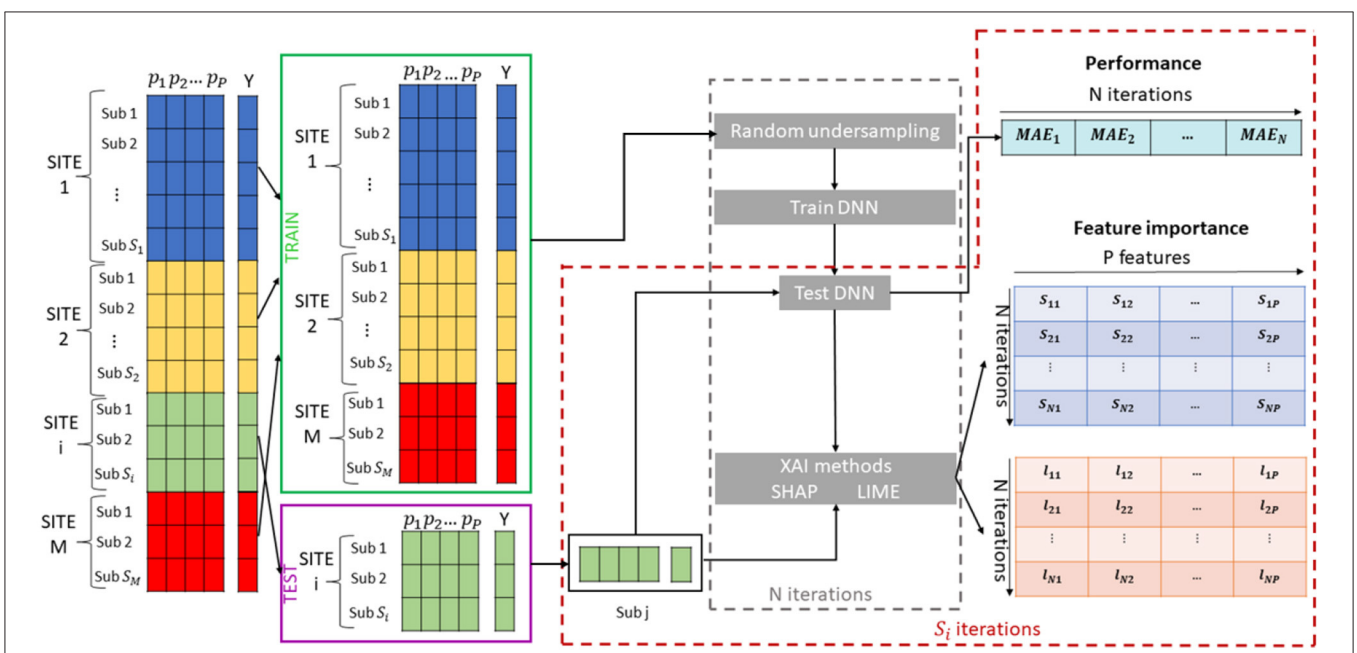

FIGURE 1 | Schematic overview of the Explainable DL framework for brain age prediction.

investigate the effect of the variability of the training set on both performance and age-related feature importance at subject-level and across subjects. In the following sections, each step of the algorithm is further explained.

3.1. Deep Neural Networks

We developed a fully connected DNN architecture. The model and all the computation was implemented using Tensorflow 2.0 (Abadi et al., 2016), with the serial interface. The Input layer shape was composed by 1, 213 units, i.e., the number of features that characterize each subject.

It is well-known that there is no general rule to determine the model hyper-parameters, so we tuned them with a series of 10-fold Grid Search cross validations on training sets, using the left out site as a completely independent test set. In each training the decrease of the loss was monitored using the Keras callback functions EarlyStop, with *patience* = 20, and ModelCheckpoint, in order to stop the training before overfitting. The parameters determined with cross validations were: the activation function (we checked ReLu and tanh), the dropout rate $(0, 0.1, 0.2, 0.3)$, the number of neurons $(128, 256, 512, 1, 024)$, the net optimizer (SGD, Adam), the loss function (Huber and MSE), the learning rate $(1e-4, 5e-5, 1e-5)$, the number of layers $(3, 4, 5)$ and the batch size $(20, 100, 400)$. We reached the final configurations with 4-layers with 512 units per layer, relu as activation function, the SGD optimizer with learning rate $5e-5$ and momentum 0.9, the loss function Huber and dropout 0. The number of epochs of each training round was controlled by the trend of the loss function on the validation subset through the callbacks mentioned above. The output layer had a single unit with no activation function, in order to perform the required regression.

The performance of the models were evaluated by means of the Mean Absolute Error (MAE):

$$MAE = \frac{1}{t}\sum_{i=1}^{t}|y_i - \hat{y}_i|, \tag{1}$$

with t being the sample size for the specific test site, y_i the chronological age, and $\hat{y}_i$ the predicted brain age. The correlation coefficient between the chronological age and the predicted age of the subjects was also computed to assess the performance of the models over the whole dataset:

$$R = \frac{\sum_{i=1}^{T}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{T}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{T}(\hat{y}_i - \bar{\hat{y}})^2}}, \tag{2}$$

where $\bar{y}$ and $\bar{\hat{y}}$ denote the sample mean of the chronological age and the predicted brain age, respectively. A non-parametric permutation test was applied to assess the statistical significance of above-chance predictive performance for the overall model as suggested in Hilger et al. (2020). In details, we permuted 1, 000 times the age outcomes of the subjects and assessed both performance values (MAE and R) within each permutation round. Finally, a p-value for each performance metric was assigned by dividing the number of times for which model performance based on the true age was lower than the performance for the permuted age outcomes by the number of permutations, i.e., 1,000.

3.2. Explainable Algorithms

In this work, we adopted the most popular local explanation algorithms: SHAP (Lundberg and Lee, 2017) and LIME (Ribeiro et al., 2016), to explain the decisions of the DNN models on each test sample. These methods are local model-agnostic as they explain predictions at individual level regardless the selected models. Basically, the two methods learn an interpretable linear model around each test instance and estimate feature importance at local level. For a dataset $D = [(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), ..., (\boldsymbol{x}_T, y_T)]$, where $\boldsymbol{x}_i \in \mathbb{R}^P$ is the feature vector for the sample i and y_i the corresponding age, the generic pre-trained model f returns a prediction $f(\boldsymbol{x}_i)$ based on a single input sample $\boldsymbol{x}_i$. SHAP and LIME aim at finding a linear model g to explain f by using a simplified inputs x' that map the original inputs through a mapping function $x = h_x(x')$ trying to ensure $g(z') \approx f(h_x(z'))$ whenever $z' \approx x'$. Both methods minimize the following objective function:

$$\xi = \underset{g \in G}{\arg\min}\, \mathcal{L}(f, g, \pi_{x'}) + \Omega(g), \tag{3}$$

where G is the class of linear models, $\pi_{x'}$ represents a proximity metric between x and x', $\Omega(g)$ denotes the complexity of the explanation g and the loss function $\mathcal{L}$ is defined as:

$$\mathcal{L}(f, g, \pi_{x'}) = \sum_{x' \in X'}[f(x') - g(x')]^2 \pi_{x'}, \tag{4}$$

where X' is the set of inputs within the neighborhood of x'. Both methods try to generate an explanation for x that approximates the behavior of the model accurately within the neighborhood of x, while achieving lower complexity (Slack et al., 2020). In other words, the methods explain the prediction of the instance x by computing the contribution of each feature to the prediction, so the absolute value of each SHAP and LIME value expresses how much each feature contributes to the final prediction (Wang et al., 2020). In LIME, $\Omega(g)$ and $\pi_{x'}$ are defined heuristically, while in SHAP they are determined by satisfying some equations from the cooperative game theory. More details about the principles underlying these methods and the mathematical definitions can be found in the seminal works of Lundberg and Lee (2017) and Ribeiro et al. (2016). In our analysis, we applied the python implementation of the SHAP[4] and LIME[5] methods.

3.3. Reliability of Explainable Scores

Both SHAP and LIME are *post-hoc* local XAI methods as they exploit a pre-trained ML model to compute approximations of the model's inner decision logic by producing understandable representations in the form of feature importance scores for each independent test sample that represent the contribution of

[4]https://anaconda.org/conda-forge/shap
[5]https://anaconda.org/conda-forge/lime

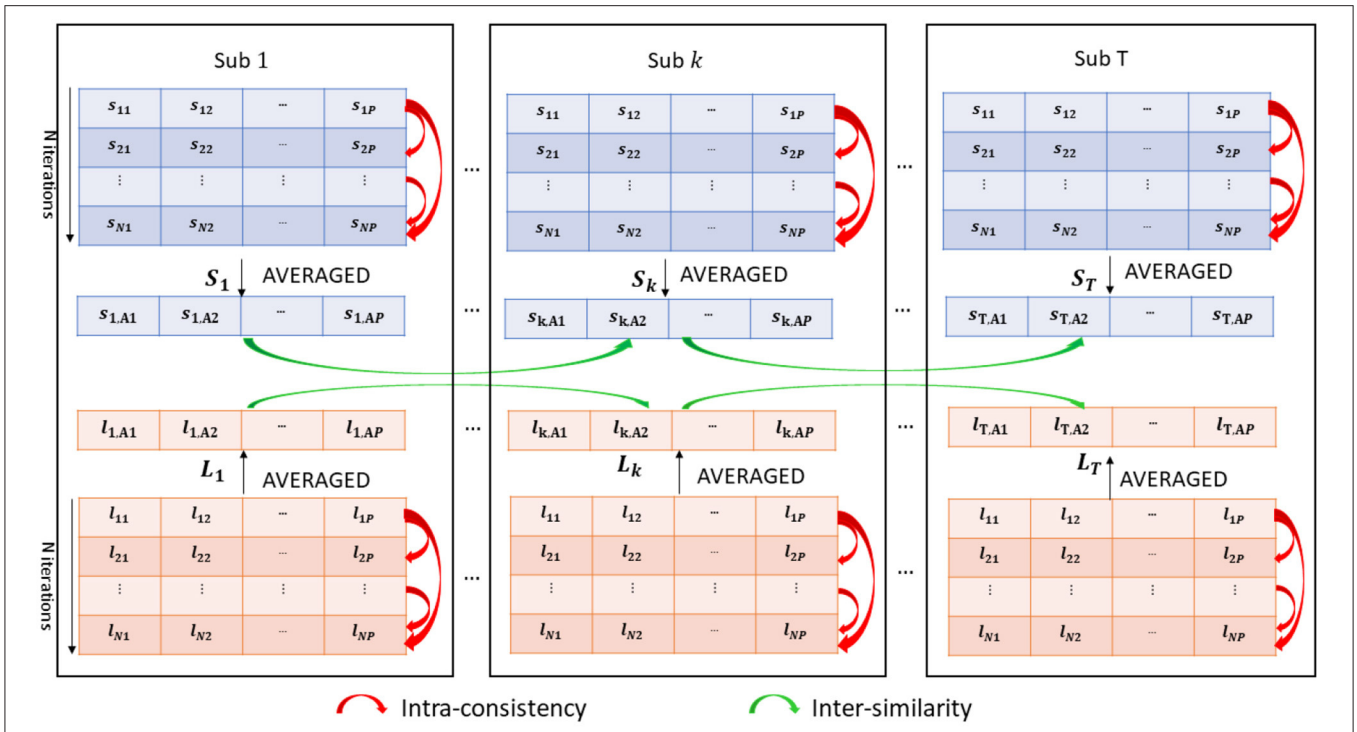

FIGURE 2 | Analysis of intra-consistency and inter-similarity of both SHAP (blue) and LIME (orange) scores.

each feature to the final prediction of the ML model (Moradi and Samwald, 2021). These methods greatly differ from feature selection methods, which use the entire train set to determine the impact of each feature on a performance metric: the output of a feature selection scheme usually results in a single feature importance vector, whereas local XAI methods output a feature importance vector for each test sample. Therefore, a reliability analysis of XAI scores was performed to quantify the variation of the score values by slightly varying the composition of the training set. Moreover, since cognitive phenotypic variables and confounding factors related to the acquisition sites can affect the morphological feature values (Frangou et al., 2004; Shaw et al., 2006; Fortin et al., 2018), we investigated whether these factors could also influence the values and the reliability of the XAI scores.

An overview of the methodology is shown in **Figure 2**. We collected $N = 100$ realizations of both SHAP and LIME vectors forming the two matrices S and L for each subject. We also averaged the $N = 100$ realizations of both values in order to obtain a single representative SHAP vector ($S_t = [s_{t,A1}, ..., s_{t,AP}]$) and LIME vector ($L_t = [l_{t,A1}, ..., l_{t,AP}]$) for each subject t, where:

$$s_{t,Ap} = \frac{1}{N}\sum_{n=1}^{N} s_{np} \tag{5}$$

is the p^{th} averaged SHAP value for the feature p.

In order to investigated the reliability of both SHAP and LIME values, we computed:

- The intra-consistency coefficient of the scores, i.e., the correlation between each couple of score vectors $\boldsymbol{s}_k = [s_{k1}, s_{k2}, ..., s_{kP}]$ and $\boldsymbol{s}_z = [s_{z1}, s_{z2}, ..., s_{zP}]$, with $k, z = 1, ..., N$ within each subject:

$$IC_{kz} = \frac{\sum_{p=1}^{P}(s_{kp} - \bar{s_k})(s_{zp} - \bar{s_z})}{\sqrt{\sum_{p=1}^{P}(s_{kp} - \bar{s_k})^2}\sqrt{\sum_{p=1}^{P}(s_{zp} - \bar{s_z})^2}}, \tag{6}$$

  where $\bar{s_k}$ and $\bar{s_z}$ denote the sample means of the two vectors and k and z denote the indices of different model training iterations. We also computed IC_{kz} for each couple of LIME vectors obtaining a distribution of $N\frac{(N-1)}{2}$ intra-consistency values for each XAI method. The intra-consistency coefficient varies between 0 (zero) and 1 (one), hence we compared the IC distributions by grouping the subjects according to their site. In addition, the correlation between the mean IC value of each subject and the variables age, FIQ and SNR was computed to verify if a possible association exists between the IC values of the subjects and each of the phenotypic information and imaging-related quality metric;
- The inter-subject similarity, i.e., the correlation between the SHAP (LIME) score vectors $\boldsymbol{S}_t$ and $\boldsymbol{S}_u$ ($\boldsymbol{L}_t$ and $\boldsymbol{L}_u$) for each couple of subjects u and t, with $t, u = 1, ..., T$:

$$IS_{ut} = \frac{\sum_{p=1}^{P}(s_{u,Ap} - \bar{S_u})(s_{t,Ap} - \bar{S_t})}{\sqrt{\sum_{p=1}^{P}(s_{u,Ap} - \bar{S_u})^2}\sqrt{\sum_{p=1}^{P}(s_{t,Ap} - \bar{S_t})^2}}, \tag{7}$$

where $\bar{S_u}$ and $\bar{S_t}$ denote the sample means of the two vectors. We then constructed an inter-similarity matrix *IS*, where entry $(u, t) = IS_{ut}$ indicates the similarity value between the scores of subjects u and t for each of the two XAI methods obtaining matrices IS_{SHAP} and IS_{LIME}. We applied the k-medoid algorithm on each *IS* matrix to find the best partition into clusters (more details on the algorithm are reported in Supplementary section 3 of **Supplementary Material**). The identified clusters of subjects were analyzed by using different criteria:

1. The site membership to investigate a possible relationship between the XAI scores and the site of the subjects;
2. The age, FIQ, and SNR distributions for comparing the phenotipic and imaging-related values across clusters. To analyze the differences between the identified clusters for each of the three variables, Kruskal–Wallis tests were conducted ($\alpha = 0.05$ with Bonferroni corrections), followed by *post-hoc* Tukey–Kramer tests in case of significant group effects.

3.4. Comparison Between SHAP and LIME

We performed a direct comparison between the SHAP and LIME scores of each subject t, by computing the correlation between the SHAP and LIME vectors $\boldsymbol{S}_t$ and $\boldsymbol{L}_T$, with $t = 1, ..., T$:

$$R_{SL,t} = \frac{\sum_{p=1}^{P}(s_{t,Ap} - \bar{S_t})(l_{t,Ap} - \bar{L_t})}{\sqrt{\sum_{p=1}^{P}(s_{t,Ap} - \bar{S_t})^2}\sqrt{\sum_{p=1}^{P}(l_{t,Ap} - \bar{L_t})^2}}, \quad (8)$$

where $\bar{S_t}$ and $\bar{L_t}$ denote the sample means of the two vectors.

In order to identify the set of morphological descriptors whose importance is most likely to vary with age, a correlation analysis was conducted between each of the SHAP and LIME averaged values and the age of the subjects. We considered $\alpha = 0.05$ with Bonferroni corrections.

We also compared the set of most significant features between the two methods to verify a possible overlap between the two sets by means of the Jaccard coefficient:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}, \quad (9)$$

where A and B are two sets of significant features resulting from SHAP and LIME, respectively. The overlapping analysis was conducted by varying the threshold level between the $75th$ and $99th$ upper percentile and $1st$ and $25th$ lower percentile of the distributions of the correlation values with step $\Delta = 2$. A non-parametric permutation test was performed by randomly permuting 1,000 times the correlation scores and assessing the percentage of overlap between the two sets to determine the statistical significance of the actual overlap for each threshold.

4. RESULTS

4.1. Performance of DNN Models

Figure 3 shows the performance of the DNN models for the different sites and for each subject. In particular, **Figure 3A** shows a bubble plot reporting information on the average MAE for each site coded by colors, the number of samples along the y axis and the average age of the subjects within each site coded by the radius of each bubble. **Figure 3B** shows the violin plots of the age distributions of the subjects within each site sorted by increasing MAE coded by the same color map of **Figure 3A**. It is clearly evident from both plots that the MAE values are related to age range within each site: the higher the age range, the higher the average MAE within a site. Notably, the sites MaxMun and SBL which include subjects with age in the last percentile of the age distribution of the whole data samples (age $>$ 30 years) resulted the sites with the worst performance highlighting a sample size effect on this age range. This finding is better explained by inspecting **Figure 3C** which shows the averaged MAE for each subject in a scatter plot reporting also the chronological age and the predicted age with their marginal distributions: the chronological age distribution is highly right-skewed and the performance of the DNN models dramatically worse in the most sparse age range, i.e., the right tail of the distribution. For the whole dataset, we found the overall performance $MAE = 2.7$ and the correlation between the chronological and predicted age of the subjects $R = 0.86$. Both metrics were found to be significantly different from the chance level, resulting $p = 0$ from the non-parametric permutation test (see **Supplementary Figure 1** for more details).

4.2. Intra-consistency of Explainable Scores

The intra-consistency coefficients of the XAI scores provide indices of consistency of the feature importance as the training set varies from round to round. **Figure 4** shows the distributions of these indices for the different sites for the SHAP scores (**Figure 4A**) and for the LIME scores (**Figure 4B**). Apart from a slight difference between the different sites for both scores, the LIME scores show consistently lower intra-consistency values (lower than 0.4 for all the sites) than those exhibited by the SHAP scores (greater than 0.5 for all the sites). Please refer to **Supplementary Figure 2** for more details. We also evaluated the correlation between the averaged intra-consistency values and each of the age, FIQ and SNR variables to investigate whether any link exists between the average intra-consistency coefficients of both XAI methods and any of the biological, phenotypic and image-related characteristics of the subjects. **Figure 5** shows the correlation between the averaged intra-consistency of the subjects and their age, FIQ and SNR for the SHAP method (**Figures 5A–C**) and LIME method (**Figures 5D–F**). Except for a weak correlation between the averaged intra-consistency indices of the SHAP values and the age of the subjects ($R = 0.10$, $P = 0.049$, not significant after Bonferroni correction), no significant correlations were observed for the other variables for both XAI methods.

4.3. Inter-similarity of Explainable Scores

We obtained two inter-similarity matrices (IS_{SHAP} and IS_{LIME}) by computing the inter-similarity coefficient between each couple of average XAI score vectors of the subjects for each XAI

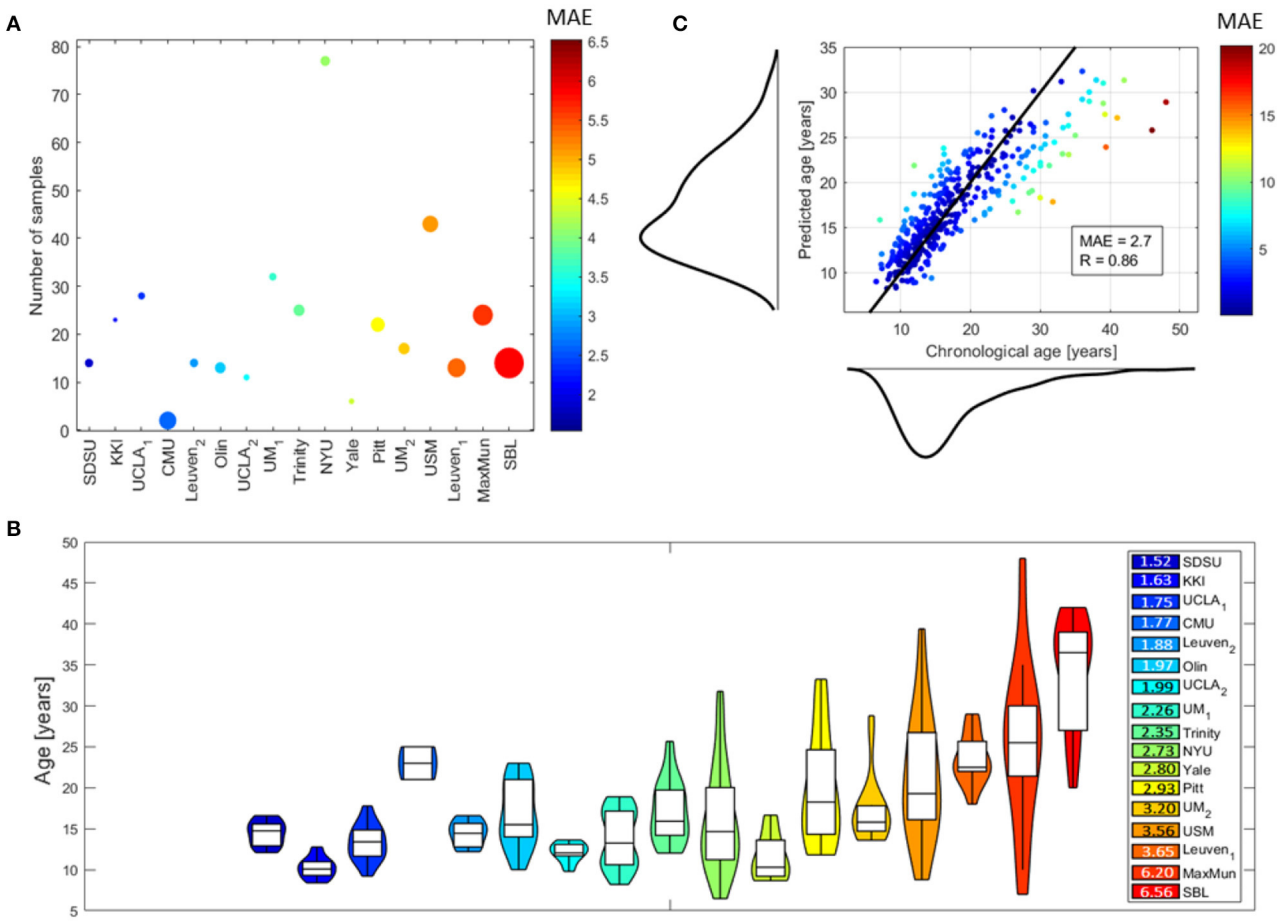

FIGURE 3 | Performance of DNN models: **(A)** bubble plot where each bubble represents a site, the color codes the average MAE values of the subjects within each site and the radius is proportional to the average age of the subjects within each site, the y position indicates the number of samples within each site; **(B)** violin plots of the age distributions of the subjects within each site (the colors indicate the average MAE values for the subjects within each site); **(C)** scatter plot showing the chronological age of the subjects vs. their predicted age with the marginal distributions (the black solid line indicates the ideal model and the color of each point codes the average MAE for the corresponding subject). The performance for the whole dataset are also reported by means of the MAE and R-values.

method. The k-medoid method was used to assess the best partition of each *IS* matrix into clusters. We found $k = 10$ for matrix IS_{SHAP} and $k = 5$ for matrix IS_{LIME}. More details on the clustering algorithm can be found in Supplementary section 3 of **Supplementary Material**. **Figure 6** shows the pie charts reporting the site membership of the subjects within each of the 10 clusters for the SHAP values and five clusters for the LIME values. The different clusters are composed of individuals from different sites, apart from cluster 2 of the IS_{SHAP} matrix containing only individuals from the NYU site and cluster 1 of matrix IS_{LIME} composed mainly of subjects from the NYU site. We analyzed also the age, FIQ and SNR distributions of the subjects within each cluster of both inter-similarity networks. **Figure 7** highlights that the age effect is greater in the IS_{SHAP} network as the age distributions of the different clusters differ more from each other (Kruskal–Wallis test: $p < 10^{-6}$, Bonferroni corrected). The imaging quality was also detected as a strong effect in the IS_{SHAP} network as the SNR distributions of the clusters are significantly different (Kruskal–Wallis test: $p < 10^{-6}$, Bonferroni corrected). More details on *post-hoc* tests are included in Supplementary section 4 of **Supplementary Material**.

4.4. Comparison Between SHAP and LIME

By directly comparing the SHAP and LIME vectors for each subject we found the average value $R_{SL} = 0.52 \pm 0.05$, showing a weak correlation value between the two XAI scores.

Figure 8 shows different results about the correlation analysis between the XAI scores and the age of the subjects:

- The distribution of coefficient values between the SHAP scores of the morphological features and the age of the subjects is significantly higher than the distribution of coefficient values between the LIME scores and the age (Wilcoxon test: $p = 10^{-6}$; Cohen's $d = 1.61$) (see **Figure 8C**);

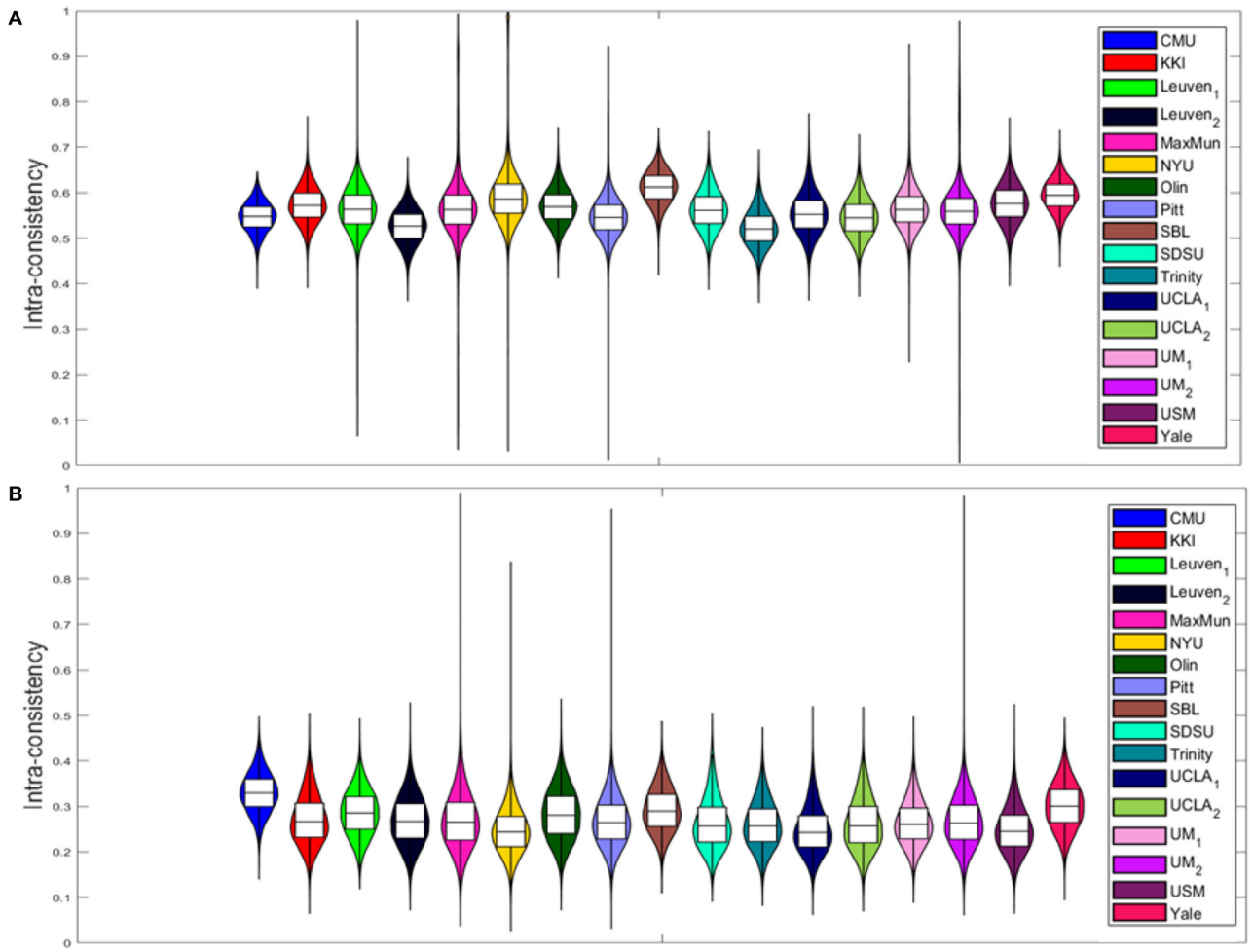

FIGURE 4 | Violin plots of the intra-consistency distributions of **(A)** SHAP values and **(B)** LIME values for each site.

- A higher number of features statistically related to the age resulted from SHAP values than from the LIME values as presented in **Figures 8A,B**, which show the Manhattan plots representing the *p*-values resulting from the correlation analysis between the XAI scores of the features and the age of the subjects;
- The sets of age-related features for the two XAI methods exhibit a remarkably low overlap. Indeed, **Figure 8D** shows that for different threshold values of the two correlation distributions, the overlap coefficient between the two sets of features is below 0.02. This point is more obvious when comparing the two sets of morphological features with the most significant correlation between the XAI scores and the chronological age (at the threshold $2 - 98$ percentiles of correlation distributions) for the two methods SHAP and LIME, listed in **Tables 2**, **3**, respectively. A significant overlap was obtained for each threshold ($p < 0.005$). The brain regions related to the two sets of features are also represented in **Figure 9**. The two sets overlap only for one feature ($p < 0.002$), i.e., the curvature index of the right precentral ROI. Moreover, among the age-related features detected with the SHAP method, a prevalence of positively age-associated cortical features is reported, whereas a prevalence of negatively age-associated volumetric WM features is observed for the LIME method.

5. DISCUSSION

In this work, we developed a novel XAI framework to perform brain age prediction with DNN and morphological features and compare the local explanations of the two XAI methods: SHAP and LIME. We adopted a cohort of healthy controls from ABIDE I dataset, whose heterogeneity is related to the different number of samples per site, the non-uniform age ranges per site and the various acquisition protocols. Hence, a leave-one-site cross validation strategy was chosen to investigate the site effect. Indeed, both the imaging and phenotipic heterogeneity of the data sample could affect the learning process and the final accuracy of the ML algorithms. The results the DNN models achieved compare favorably with the literature showing the

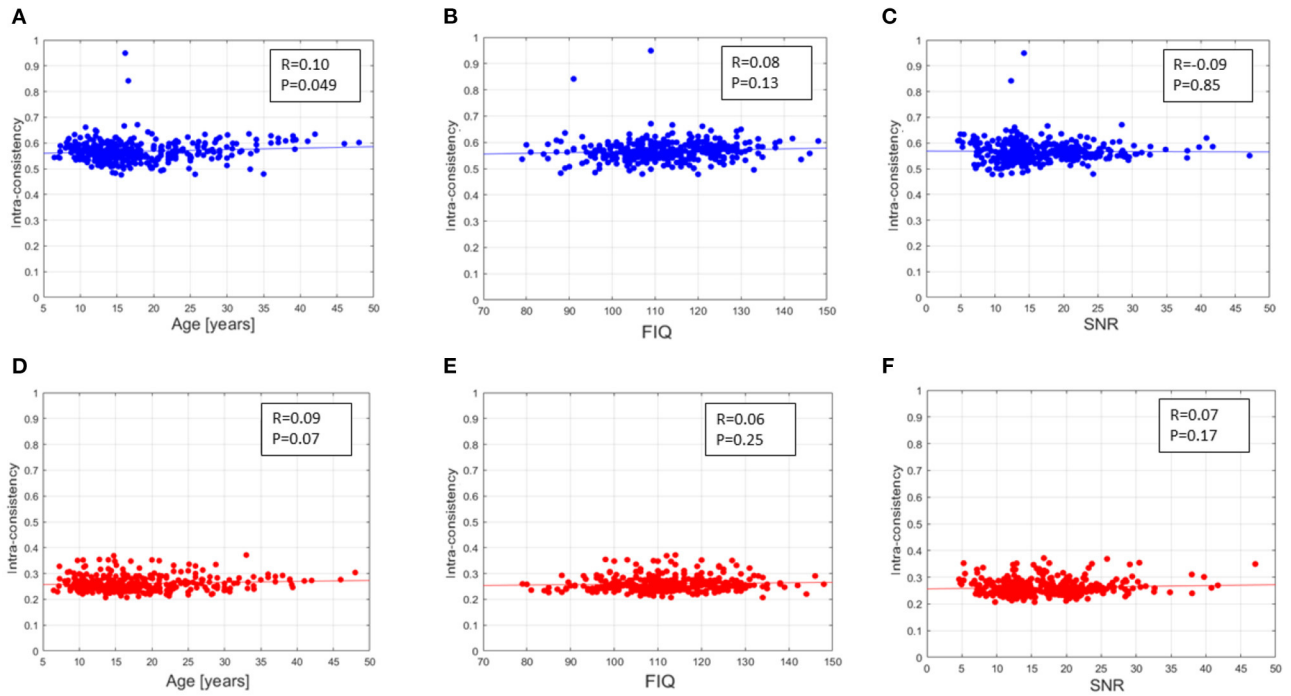

FIGURE 5 | Correlation found between the averaged intra-consistency of SHAP values and **(A)** age, **(B)** FIQ and **(C)** SNR of the MRI of each subject; correlation found between the averaged intra-consistency of LIME values and **(D)** age, **(E)** FIQ and **(F)** SNR of the MRI of each subject.

overall performance $MAE = 2.7$ and $R = 0.86$ (Ball et al., 2019; Corps and Rekik, 2019,?; Zhao et al., 2019; Bellantuono et al., 2020). However, as shown in **Figure 3**, our models exhibit a systematic age under-estimation in the most extreme age-range of the distribution, reporting worse performance ($MAE > 4$) at sites with individuals with chronological age in that range. We found similar results in our previous work in which more simple ML models were applied on the same dataset (Lombardi et al., 2020b).

Afterwards, we applied the two XAI methods to each sample to derive the local explanations, i.e., a feature importance vector that express the contributions of the morphological features to the final prediction of the DNN models. We performed a hierarchical analysis to compare the reliability of the XAI scores both at subject-level and across subjects. Firstly, an intra-consistency score was defined to objectively assess the stability of the XAI methods for each subject with respect small variations of the training set. Indeed, the feature importance at local level should not vary significantly by slightly perturbing the composition of the training set in order to define a reliable personalized final ranking of the morphological features for the age prediction task (Kalousis et al., 2005; Lombardi et al., 2020a). We compared the intra-consistency values of both SHAP and LIME scores across the sites to verify a possible site effect on the feature importance. **Figure 5**, **Supplementary Figure 2** and **Supplementary Table 1** clearly highlight that some significant differences in intra-consistency values exist between some sites for both methods. Moreover, it is worth noting that the two XAI methods exhibit very different intra-consistency values as the SHAP intra-consistency scores are significantly higher than the LIME scores regardless the acquisition site. We computed the correlation coefficient between the averaged intra-consistency values of the subjects and each of the variables age, FIQ and SNR for both XAI methods to investigate the relationship between the feature importance and the phenotipic and imaging-related information. **Figure 5** outlines that none of this variables is significantly related to the intra-consistency of both SHAP and LIME scores, hence the reliability of the XAI scores at local level does not depend on these characteristics of the subjects.

In order to compare the XAI scores across the subjects, we defined an inter-similarity score. We computed a single vector of SHAP and LIME scores for each subject by averaging the vectors resulting from the N under-sampling rounds of the training set. The purpose of this step was to obtain a single consistent feature importance vector for each method as by averaging the different realizations, the more stable scores are enhanced, while the more fluctuating scores are de-emphasized. We correlated the XAI vectors between each couple of subjects to assess the similarity of XAI scores among the subjects, i.e., the inter-similarity score. Finally, an inter-similarity matrix IS was constructed for each method and partitioned into clusters to detect groups of subjects with similar XAI scores. The analysis of the detected clusters show that except for a single cluster composed only by subjects from site NYU for the SHAP method and a cluster composed mainly of subjects from the NYU site for the LIME method, the other clusters include subjects from

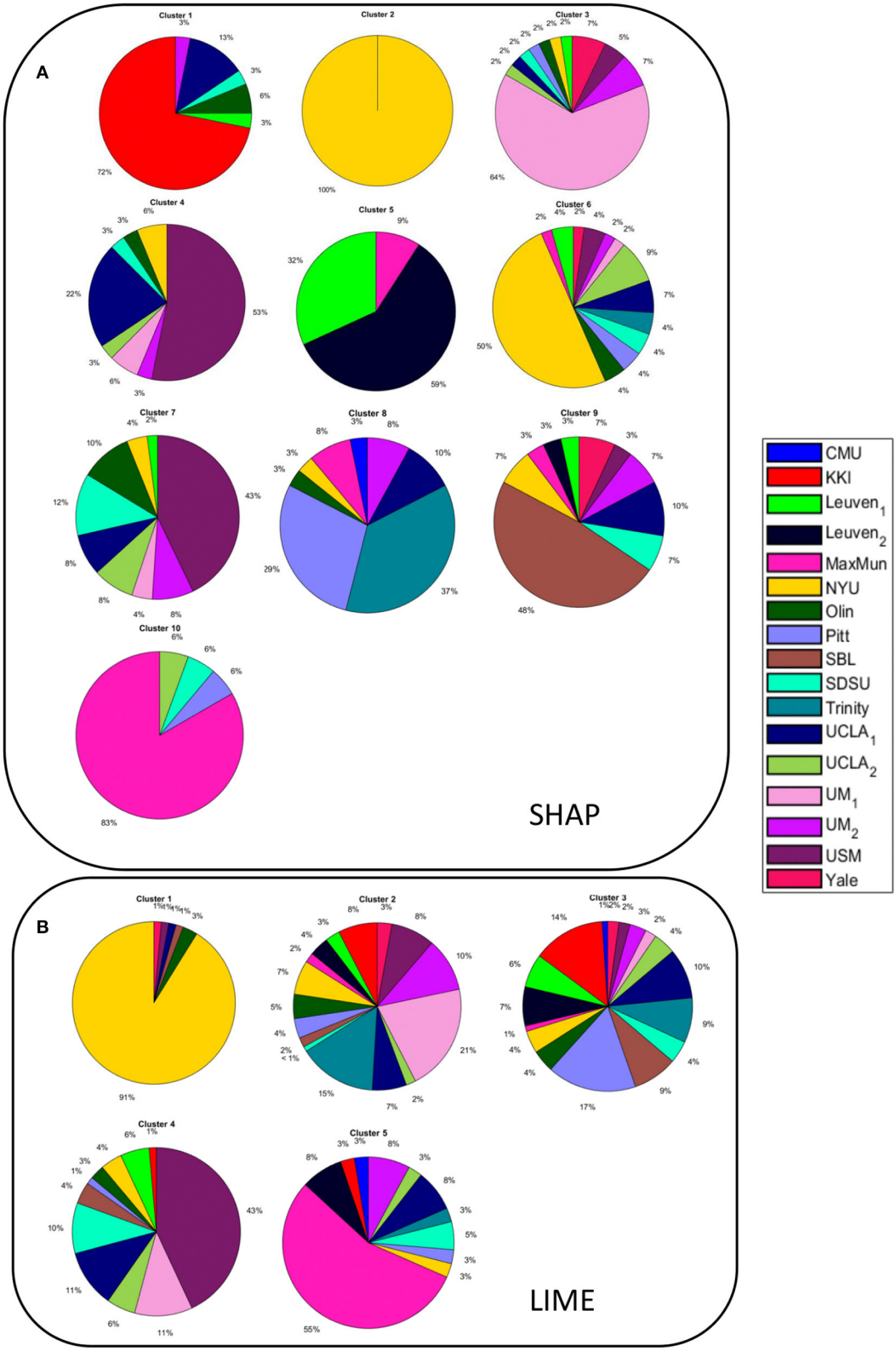

FIGURE 6 | Pie charts showing the site membership of **(A)** the subjects belonging to each of the 10 clusters resulting from the partition of the inter-similarity network of the SHAP values; **(B)** the subjects belonging to each of the 5 clusters resulting from the partition of the inter-similarity network of the LIME values.

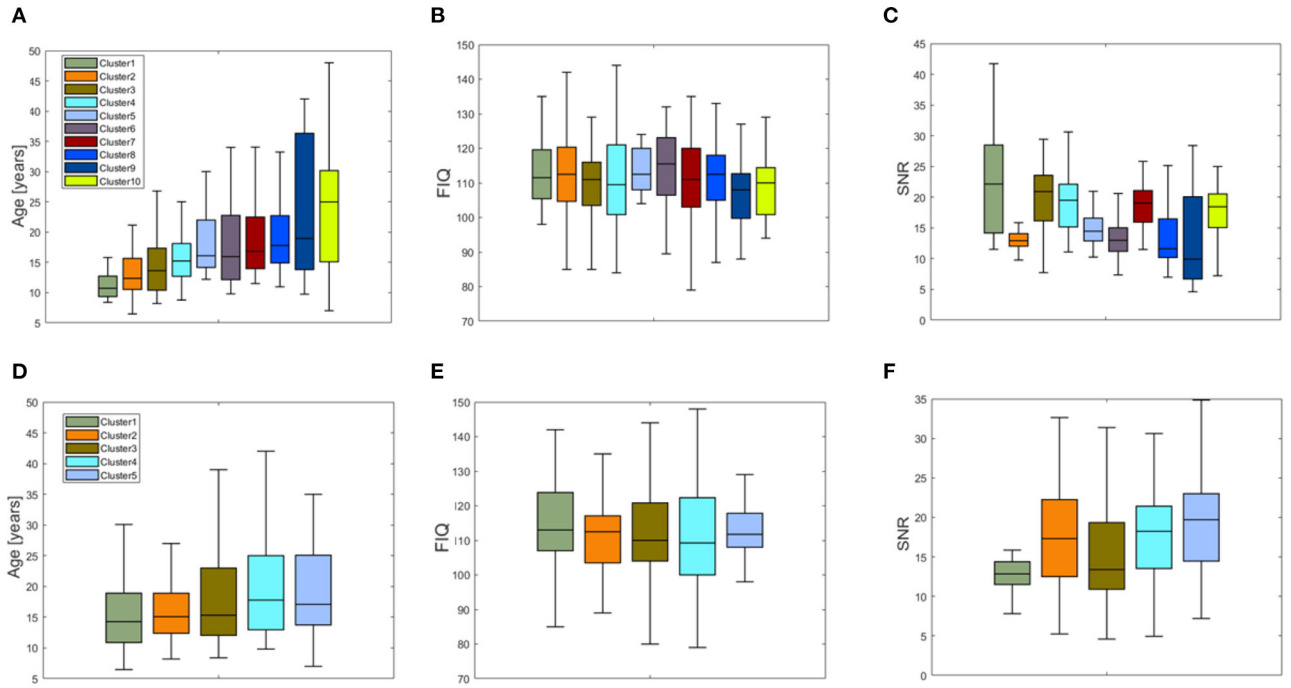

FIGURE 7 | Boxplots of the distribution of **(A)** the age, **(B)** the FIQ values, and **(C)** SNR values for the 10 clusters resulting from the partition of the inter-similarity network of the SHAP values; boxplots of the distributions of **(D)** the age, **(E)** the FIQ values, and **(F)** SNR values for the five clusters resulting from the partition of the inter-similarity network of the LIME values.

different sites (see **Figure 6**). This finding indicates that the feature importance values extracted by the two XAI methods also reflect the different characteristics of the NYU site with respect to all the other sites that have been discussed in several studies (Shehzad et al., 2015; Bhaumik et al., 2018). Similarly to the subject-based analysis, we compared the phenotypic and imaging quality-related variables between the different clusters for the two XAI methods. From **Figure 7**, it can be noted that only the clusters derived from the IS_{SHAP} matrix reflect a partition of the subjects into significantly different age ranges, whereas the clusters extracted from the IS_{LIME} matrix do not reveal a clear age-related partition of the subjects. This finding confirms the reliability of the SHAP scores with respect to the age prediction problem. Moreover, both partitions are related to the SNR of the subjects as the clusters also significantly differ for the imaging-related quality metric, showing that the site heterogeneity also affects the local XAI scores as well as the performance of the predictive models.

We performed a direct comparison between the SHAP and LIME vectors which highlighted a low correlation between the XAI scores for each subject. Moreover, a correlation analysis between each feature score vector and the age of the subjects was performed to yield a set of morphometric descriptors whose relevance for age prediction is most variable with age. This step of the framework provides global explanations of the DNN models since a set of age-related scores is extracted from the whole population under investigation. As shown in **Figure 8C**, the correlation distributions between the XAI scores and the age are markedly different from each other ($p < 10^{-6}$, Cohen's $d = 1.61$). We reported the most age-related features for SHAP and LIME methods at the statistical threshold of the 97^{th} percentile of the correlation distributions in **Tables 2**, **3**, respectively. The brain regions corresponding to the most age-related features for both XAI methods are shown in **Figure 9**. Average thickness, folding, and curvature index statistical attributes related to precentral gyrus and inferior and lateral occipital cortex were detected as the most age correlated for the SHAP method. Relevant morphological changes of these regions have been reported both in neurodevelopment and aging trajectories (Tamnes et al., 2010; McGinnis et al., 2011; Remer et al., 2017). In addition, changes in cortical curvature and folding of these regions have been extensively observed during brain maturation (Meng et al., 2014; Lefèvre et al., 2015). We also found CSF statistical descriptors as features significantly correlated with age in line with several works where cerebrospinal fluid biomarkers have been identified for normal aging process as well as for brain atrophy characterization (Preul et al., 2006; Baird et al., 2012; Vinke et al., 2018). In contrast, these regions do not appear among the most age-related LIME scores. In this set, features related to WM volumes of opercular and triangular part of inferior frontal gyrus and inferior temporal gyrus were detected as the most age-related descriptors. Notably, only the SHAP method showed a significant correlation between the importance of the cortical thickness of both hemispheres and

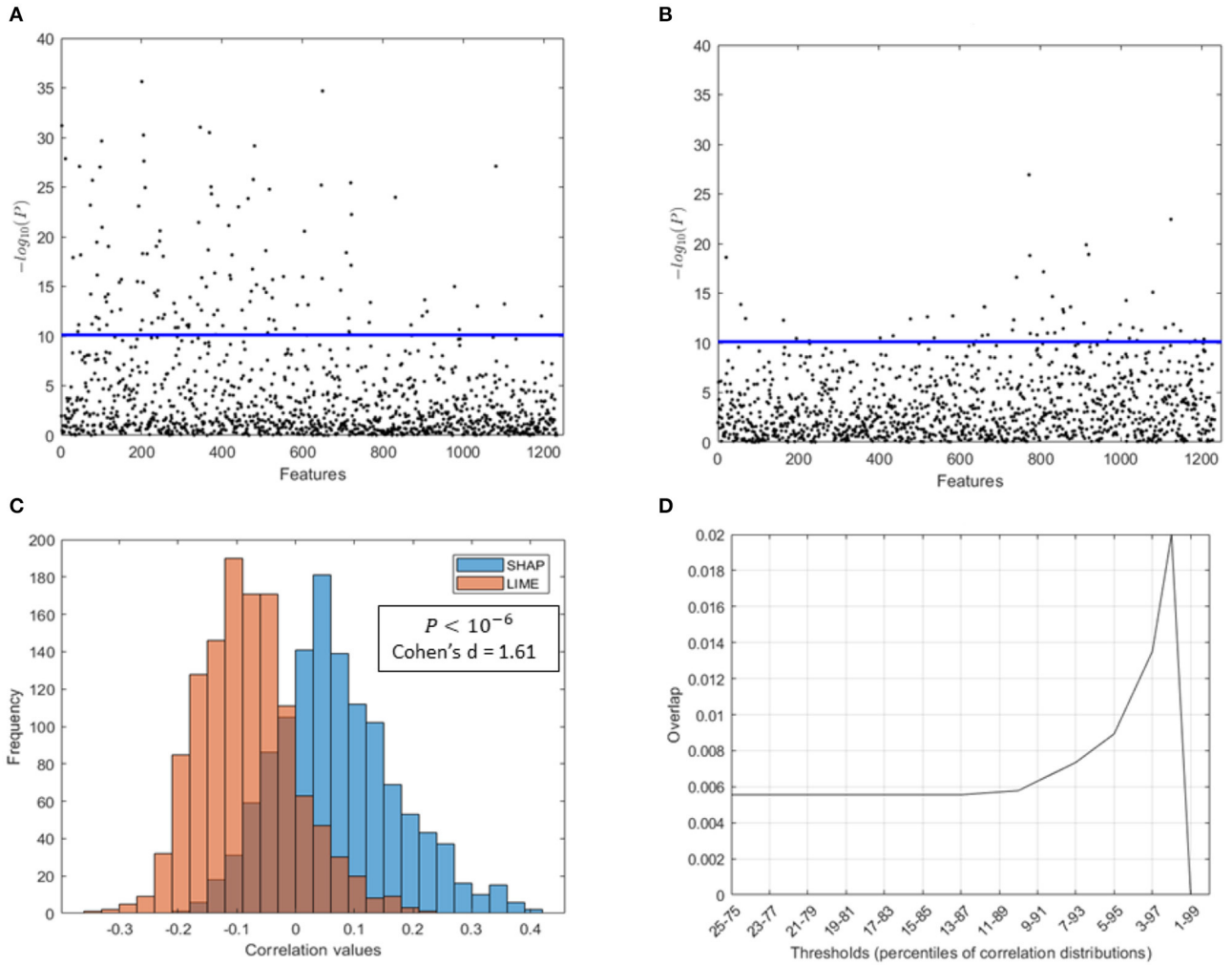

FIGURE 8 | Comparison of SHAP and LIME scores. Manhattan plot representing the p values resulting from the correlation analysis between **(A)** the SHAP scores of the features and the age of the subjects, **(B)** the LIME scores of the features and the age of the subjects; **(C)** distributions of correlation values found between the SHAP/LIME scores of the features and the age of the subjects; **(D)** overlap between the most significant features for the two criteria (SHAP/LIME) selected by varying the percentile threshold of the correlation distributions.

age ($R = 0.38$ for left and $R = 0.36$ for right). This finding is highly consistent with several previous studies which, although reporting non-linear and widespread regional variations of both cortical and volume morphology with age, unequivocally agree on cortical thinning as a pattern of neurodevelopment (Zielinski et al., 2014; Fjell et al., 2015; Tamnes et al., 2017). In general, the age-related feature sets for the two methods strongly differ from each other as shown in **Figure 8**. Indeed, the overlapping analysis between the two sets highlights a low overlap, regardless the selected correlation threshold. In addition, a prevalence of negative age correlation values can be observed for the LIME scores. A significant negative correlation between the LIME values of a given morphological feature and the age of the subjects means that the LIME importance of that feature decreases as age increases. However, it is not possible to claim that the LIME scores better explain age in younger subjects than in older subjects as in our analysis we found that the LIME scores showed very low intra-consistency values regardless of the age of the subjects (as shown in **Figure 5D**).

6. LIMITATIONS AND FUTURE DIRECTIONS

Although this work shows some important implications of SHAP and LIME XAI methods for the interpretations of brain age predictions with DNN models, it presents some limitations. We selected a cohort of typically neurodevelopment subjects from the ABIDE I dataset to explore the effect of heterogeneity of acquisition protocols and dataset composition on XAI scores. Our analysis revealed that the site effect also influences the XAI scores and therefore upstream harmonization techniques

TABLE 2 | The most significant age-related morphological features resulting from the SHAP scores grouped by category (R, Right; L, Left; curv, mean curvature; thick, thickness; vol, volume; the correlation coefficient for each feature is reported in brackets).

Sub-cortical volume	Cortical features	WM volumes	Global features
(+0.34) CSF normStdDev	(+0.35) L caudalmiddlefrontal CurvInd	(+0.33) wm L cuneus Vol	(+0.38) L mean thick
(+0.39) CSF normRange	(+0.33) L inferiorparietal ThickAvg	(+0.35) wm Rlateralorbitofrontal normMax	(+0.36) R mean thick
(+0.34) Right Pallidum normMin	(+0.34) L inferiorparietal CurvInd		
	(+0.35) L lateraloccipital ThickAvg		
	(+0.37) L lateraloccipital FoldInd		
	(+0.40) L precentral ThickAvg		
	(+0.37) L precentral FoldInd		
	(+0.35) L precentral CurvInd		
	(+0.34) L precuneus ThickAvg		
	(+0.38) R inferiorparietal ThickStd		
	(+0.37) R lateraloccipital ThickAvg		
	(+0.34) R lateraloccipital FoldInd		
	(+0.33) R lateraloccipital CurvInd		
	(+0.32) R lingual CurvInd		
	(+0.33) R posteriorcingulate ThickAvg		
	(+0.33) R precentral CurvInd		
	(+0.36) R precuneus ThickAvg		
	(+0.34) R superiorparietal CurvInd		

TABLE 3 | The most significant age-related morphological features resulting from the LIME scores grouped by category (R, Right; L, Left; curv, curvature; thick, thickness; vol, volume; the correlation coefficient for each feature is reported in brackets).

Sub-cortical volume	Cortical features	WM volumes
(−0.23) L Cerebellum WM normMax	(−0.29) L entorhinal ThickAvg	(−0.25) L caudalmiddlefrontal normRange
(−0.25) L VentralDC normMax	(−0.25) L fusiform FoldInd	(−0.24) L inferiortemporal normMean
(−0.25) L VentralDC normRange	(−0.23) L parsorbitalis MeanCurv	(−0.24) L inferiortemporal normMin
(−0.24) R Amygdala normRange	(0.23) R precentral CurvInd	(−0.24) L lateralorbitofrontal normStdDev
(−0.27) R VentralDC Vol	(−0.23) R superiortemporal GrayVol	(−0.23) L paracentral normRange
(−0.35) non-WM hypointensities normMean		(−0.30) L parsopercularis normRange
(−0.30) non-WM hypointensities normMin		(−0.29) L parsorbitalis normRange
(−0.23) CC Mid Anterior normMax		(−0.25) L insula normStdDev
(−0.28) CC Anterior normMean		(−0.26) R lateralorbitofrontal normStdDev
		(−0.32) R parsorbitalis normRange
		(−0.22) R parstriangularis normRange

should be applied to the morphological features to reduce the batch effects (Fortin et al., 2018). Another important aspect concerning the selected cohort is its sample size and age range, indeed in our study morphological features are analyzed to predict the age of 378 subjects in the limited age range 6–48. Previous works have widely demonstrated that both the sample size and the age range could affect the performance of age prediction models (Amoroso et al., 2018; Peng et al., 2021). Moreover, currently the best state-of-the-art results have been achieved with datasets larger than 2,000 samples (Levakov et al., 2020; Peng et al., 2021). The reliability of the XAI values is closely related to the accuracy of the predictive models, so future developments will focus on training predictive models on larger cohorts with broader age range to extend the validity of our findings.

In this work we exploited a feature-based DNN age regression approach, therefore, we adopted SHAP and LIME to produce feature relevance morphological vectors as these two algorithms represent the two most established local model-agnostic XAI techniques. However, different XAI techniques have been developed to quantify the interpretability of the latent representations of CNNs: layer-wise Relevance Propagation (LRP) technique, saliency maps, and Gradient-weighted Class Activation Mapping (Grad-CAM) can be potentially used to produce coarse localization maps (Selvaraju et al., 2017; Eitel et al., 2019; Arrieta et al., 2020), highlighting the important regions in each MRI scan by exploiting the information at voxel level.

Finally, it is important to note that we performed a correlation analysis to identify the morphological descriptors

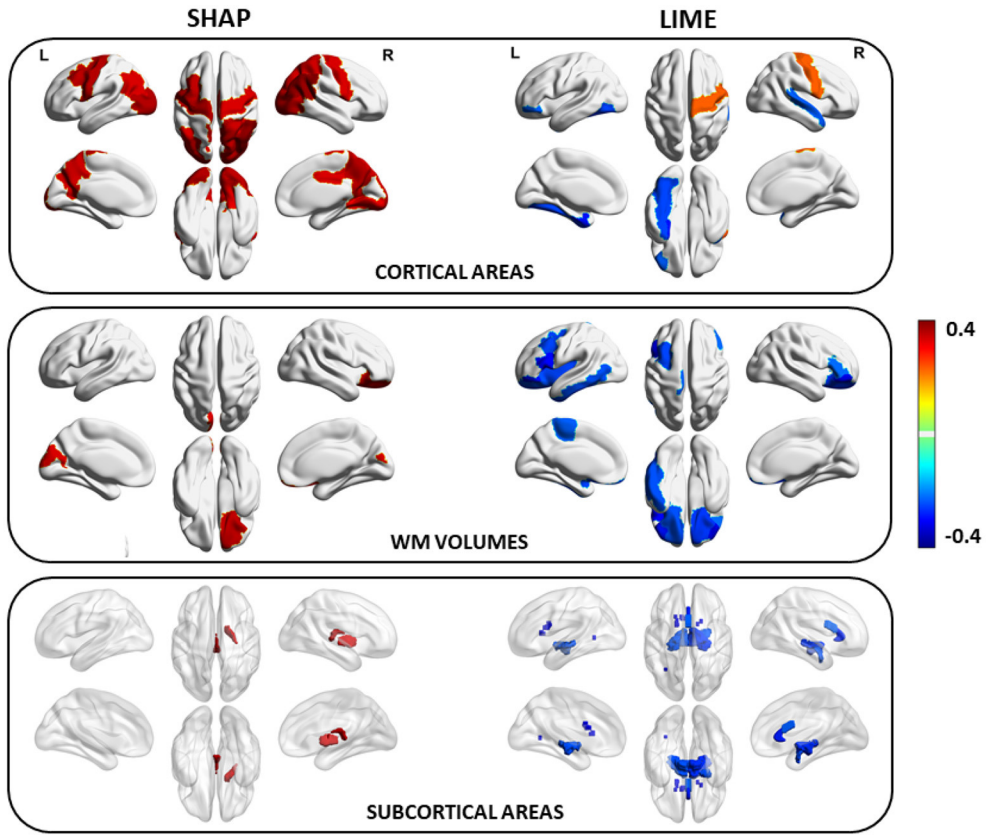

FIGURE 9 | Brain regions resulting from the morphological features with the most significant correlation between the XAI scores and the chronological age for SHAP and LIME (the color of each ROI codes the average correlation value between the XAI scores of the features related to that ROI and the age of the subjects).

whose importance most significantly varies with age. Hence, we compared the set of descriptors with the most significant correlation between the XAI scores and the age of the subjects to assess the overlap between the two XAI methods. However, a targeted and quantitative analysis is needed to compare the regions with significant impact on age prediction with the age-related regions reported in other studies. In future work, we will address a deeper comparison between the XAI scores of subjects grouped by age to existing meta-analysis.

7. CONCLUSION

In this work, we proposed a novel XAI framework to provide accurate explanations of both performance of DL algorithms and feature importance at subject level for age prediction with brain morphology. We extensively evaluated the reliability of the two XAI methods for the age prediction task both at subject level by assessing the intra-consistency of the XAI scores and across subjects by analyzing the inter-similarity of the scores. Our results reveal that the SHAP values showed significantly higher intra-consistency values than the LIME scores. This finding highlights that the SHAP scores are less influenced by small variations of the training set showing greater consistency of their values by varying the composition of the training set. Another interesting result concerns the analysis of the inter-similarity of the XAI scores between the subjects, which showed that the SHAP values more consistently reflect a partition of the subjects into different age ranges, proving therefore, a higher reliability of the SHAP scores for the age prediction task. The correlation analysis between the feature importance values and the age of the subjects showed that the two XAI methods detect totally different age-related features. In particular, the SHAP values exhibited a greater number of features statistically associated with age with higher absolute correlation values than those shown by the LIME method. Our findings indicate that the SHAP method could provide more reliable explanations for the morphological aging mechanisms that could be also exploited to identify personalized age-related imaging biomarkers.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://fcon_1000.projects.nitrc.org/indi/abide.

AUTHOR CONTRIBUTIONS

AL and ST conceived the analysis. AL performed the data curation and performed the analysis. AL, DD, ST, and RB defined the methodology. AL and DD implemented the software

pipelines and wrote the original draft. RB and ST supervised the analysis. AL, DD, ST, RB, NA, AM, and JT analyzed and interpreted the results and edited the final version of the manuscript. All authors have approved the final version of the manuscript.

FUNDING

This work was supported in part by the research project Biomarcatori di connettività cerebrale da imaging multimodale per la diagnosi precoce e stadiazione personalizzata di malattie neurodegenerative con metodi avanzati di intelligenza artificiale in ambiente di calcolo distribuito (project code 928A7C98) within the Program Research for Innovation - REFIN funded by Regione Puglia (Italy) in the framework of the POR Puglia FESR FSE 2014-2020 - Asse X - Azione 10.4.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2021.674055/full#supplementary-material

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). "Tensorflow: a system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)* (Savannah, GA), 265–283.

Abraham, A., Milham, M. P., Di Martino, A., Craddock, R. C., Samaras, D., Thirion, B., et al. (2017). Deriving reproducible biomarkers from multi-site resting-state data: an autism-based example. *Neuroimage* 147, 736–745. doi: 10.1016/j.neuroimage.2016.10.045

Adadi, A., and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* 6, 52138–52160. doi: 10.1109/ACCESS.2018.2870052

Amoroso, N., Diacono, D., Fanizzi, A., La Rocca, M., Monaco, A., Lombardi, A., et al. (2018). Deep learning reveals Alzheimer's disease onset in mci subjects: results from an international challenge. *J. Neurosci. Methods* 302, 3–9. doi: 10.1016/j.jneumeth.2017.12.011

Amoroso, N., Rocca, M. L., Bellantuono, L., Diacono, D., Fanizzi, A., Lella, E., et al. (2019). Deep learning and multiplex networks for accurate modeling of brain age. *Front. Aging Neurosci.* 11:115. doi: 10.3389/fnagi.2019.00115

Arrieta, A. B., Diaz-Rodriguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inform. Fusion* 58, 82–115. doi: 10.1016/j.inffus.2019.12.012

Baird, G. S., Nelson, S. K., Keeney, T. R., Stewart, A., Williams, S., Kraemer, S., et al. (2012). Age-dependent changes in the cerebrospinal fluid proteome by slow off-rate modified aptamer array. *Am. J. Pathol.* 180, 446–456. doi: 10.1016/j.ajpath.2011.10.024

Ball, G., Beare, R., and Seal, M. L. (2019). Charting shared developmental trajectories of cortical thickness and structural connectivity in childhood and adolescence. *Hum. Brain Mapp.* 40, 4630–4644. doi: 10.1002/hbm.24726

Bellantuono, L., Marzano, L., La Rocca, M., Duncan, D., Lombardi, A., Maggipinto, T., et al. (2020). Predicting brain age with complex networks: from adolescence to adulthood. *Neuroimage* 225:117458. doi: 10.1016/j.neuroimage.2020.117458

Bhaumik, R., Pradhan, A., Das, S., and Bhaumik, D. K. (2018). Predicting autism spectrum disorder using domain-adaptive cross-site evaluation. *Neuroinformatics* 16, 197–205. doi: 10.1007/s12021-018-9366-0

Biecek, P. (2018). Dalex: explainers for complex predictive models in R. *J. Mach. Learn. Res.* 19, 3245–3249. Available online at: http://jmlr.org/papers/v19/18-416.html

Boedhoe, P. S., Van Rooij, D., Hoogman, M., Twisk, J. W., Schmaal, L., Abe, Y., et al. (2020). Subcortical brain volume, regional cortical thickness, and cortical surface area across disorders: findings from the enigma adhd, asd, and ocd working groups. *Am. J. Psychiatry* 177, 834–843. doi: 10.1176/appi.ajp.2020.19030331

Bussmann, N., Giudici, P., Marinelli, D., and Papenbrock, J. (2020). Explainable AI in fintech risk management. *Front. Artif. Intell.* 3:26. doi: 10.3389/frai.2020.00026

Carvalho, D. V., Pereira, E. M., and Cardoso, J. S. (2019). Machine learning interpretability: a survey on methods and metrics. *Electronics* 8:832. doi: 10.3390/electronics8080832

Cole, J. H. (2017a). Neuroimaging-derived brain-age: an ageing biomarker? *Aging* 9, 1861–1862. doi: 10.18632/aging.101286

Cole, J. H., and Franke, K. (2017b). Predicting age using neuroimaging: Innovative brain ageing biomarkers. *Trends Neurosci.* 40, 681–690. doi: 10.1016/j.tins.2017.10.001

Cole, J. H., Leech, R., Sharp, D. J., and Alzheimer's Disease Neuroimaging Initiative (2015). Prediction of brain age suggests accelerated atrophy after traumatic brain injury. *Ann. Neurol.* 77, 571–581. doi: 10.1002/ana.24367

Cole, J. H., Marioni, R. E., Harris, S. E., and Deary, I. J. (2019). Brain age and other bodily "ages": implications for neuropsychiatry. *Mol. Psychiatry* 24, 266–281. doi: 10.1038/s41380-018-0098-1

Cole, J. H., Poudel, R. P. K., Tsagkrasoulis, D., Caan, 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. doi: 10.1016/j.neuroimage.2017.07.059

Corps, J., and Rekik, I. (2019). Morphological brain age prediction using multi-view brain networks derived from cortical morphology in healthy and disordered participants. *Sci. Rep.* 9:9676. doi: 10.1038/s41598-019-46145-4

Couvy-Duchesne, B., Faouzi, J., Martin, B., Thibeau-Sutre, E., Wild, A., Ansart, M., et al. (2020). Ensemble learning of convolutional neural network, support vector machine, and best linear unbiased predictor for brain age prediction: ARAMIS contribution to the predictive analytics competition 2019 challenge. *Front. Psychiatry* 11:593336. doi: 10.3389/fpsyt.2020.593336

Craddock, C., Benhajali, Y., Chu, C., Chouinard, F., Evans, A., Jakab, A., et al. (2013). The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. *Front. Neuroinform.* 7:41. doi: 10.3389/conf.fninf.2013.09.00041

Da Costa, P. F., Dafflon, J., and Pinaya, W. H. (2020). Brain-age prediction using shallow machine learning: predictive analytics competition 2019. *Front. Psychiatry* 11:604478. doi: 10.3389/fpsyt.2020.604478

Dale, A., Fischl, B., and Sereno, M. (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. *Neuroimage* 9, 179–194. doi: 10.1006/nimg.1998.0395

Desikan, R., Ségonne, F., Fischl, B., Quinn, B., Dickerson, B., Blacker, D., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* 31, 968–980. doi: 10.1016/j.neuroimage.2006.01.021

Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., et al. (2014). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* 19, 659–667. doi: 10.1038/mp.2013.78

Eitel, F., Soehler, E., Bellmann-Strobl, J., Brandt, A. U., Ruprecht, K., Giess, R. M., et al. (2019). Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation. *Neuroimage Clin.* 24:102003. doi: 10.1016/j.nicl.2019.102003

Erus, G., Battapady, H., Satterthwaite, T. D., Hakonarson, H., Gur, R. E., Davatzikos, C., et al. (2015). Imaging patterns of brain development and their relationship to cognition. *Cereb. Cortex* 25, 1676–1684. doi: 10.1093/cercor/bht425

Fellous, J.-M., Sapiro, G., Rossi, A., Mayberg, H., and Ferrante, M. (2019). Explainable artificial intelligence for neuroscience: behavioral

neurostimulation. *Front. Neurosci.* 13:1346. doi: 10.3389/fnins.2019.01346

Feng, X., Lipton, Z. C., Yang, J., Small, S. A., Provenzano, F. A., Alzheimer's Disease Neuroimaging Initiative, et al. (2020). Estimating brain age based on a uniform healthy population with deep learning and structural MRI. *Neurobiol. Aging* 91, 15–25. doi: 10.1016/j.neurobiolaging.2020.02.009

Fischl, B., Salat, D., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., et al. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron* 33, 341–355. doi: 10.1016/S0896-6273(02)00569-X

Fischl, B., Sereno, M., and Dale, A. M. (1999). Cortical surface-based analysis: Ii: inflation, flattening, and a surface-based coordinate system. *Neuroimage* 9, 195–207. doi: 10.1006/nimg.1998.0396

Fjell, A. M., Grydeland, H., Krogsrud, S. K., Amlien, I., Rohani, D. A., Ferschmann, L., et al. (2015). Development and aging of cortical thickness correspond to genetic organization patterns. *Proc. Natl. Acad. Sci. U.S.A.* 112, 15462–15467. doi: 10.1073/pnas.1508831112

Fortin, J.-P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., et al. (2018). Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage* 167, 104–120. doi: 10.1016/j.neuroimage.2017.11.024

Frangou, S., Chitins, X., and Williams, S. C. (2004). Mapping iq and gray matter density in healthy young people. *Neuroimage* 23, 800–805. doi: 10.1016/j.neuroimage.2004.05.027

Franke, K., and Gaser, C. (2012). Longitudinal changes in individual brainage in healthy aging, mild cognitive impairment, and Alzheimer's disease. *GeroPsych* 25:235. doi: 10.1024/1662-9647/a000074

Gaser, J., Franke, K., Kloppel, S., Koutsouleris, N., and Sauer, H. (2013). Brainage in mild cognitive impaired patients: predicting the conversion to Alzheimer's disease. *PLoS ONE* 8:e67346. doi: 10.1371/journal.pone.0067346

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Comput. Surv.* 51, 1–42. doi: 10.1145/3236009

Gunning, D. (2017). *Explainable Artificial Intelligence (XAI)*. Defense Advanced Research Projects Agency (DARPA).

Han, L. K., Dinga, R., Hahn, T., Ching, C. R., Eyler, L. T., Aftanas, L., et al. (2020). Brain aging in major depressive disorder: results from the enigma major depressive disorder working group. *Mol. Psychiatry* 1–16. doi: 10.1038/s41380-020-0754-0

Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., and Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the abide dataset. *Neuroimage Clin.* 17, 16–23. doi: 10.1016/j.nicl.2017.08.017

Hilger, K., Winter, N. R., Leenings, R., Sassenhagen, J., Hahn, T., Basten, U., et al. (2020). Predicting intelligence from brain gray matter volume. *Brain Struct. Funct.* 225, 2111–2129. doi: 10.1007/s00429-020-02113-7

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., and Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscipl. Rev. Data Mining Knowled. Discov.* 9:e1312. doi: 10.1002/widm.1312

Kalousis, A., Prados, J., and Hilario, M. (2005). "Stability of feature selection algorithms," in *Fifth IEEE International Conference on Data Mining (ICDM'05)* (Houston, TX: IEEE), 8. doi: 10.1109/ICDM.2005.135

Koutsouleris, N., Davatzikos, C., Borgwardt, S., Gaser, C., Bottlender, R., Frodl, T., et al. (2014). Accelerated brain aging in schizophrenia and beyond: a neuroanatomical marker of psychiatric disorders. *Schizophr. Bull.* 40, 1140–1153. doi: 10.1093/schbul/sbt142

Langlotz, C. P., Allen, B., Erickson, B. J., Kalpathy-Cramer, J., Bigelow, K., Cook, T. S., et al. (2019). A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 NIH/RSNA/ACR/the Academy Workshop. *Radiology* 291, 781–791. doi: 10.1148/radiol.2019190613

Lefévre, J., Germanaud, D., Dubois, J., Rousseau, F., de Macedo Santos, I., Angleys, H., et al. (2015). Are developmental trajectories of cortical folding comparable between cross-sectional datasets of fetuses and preterm newborns? *Cereb. Cortex* 26, 3023–3035. doi: 10.1093/cercor/bhv123

Levakov, G., Rosenthal, G., Shelef, I., Raviv, T. R., and Avidan, G. (2020). From a deep learning model back to the brain-identifying regional predictors and their relation to aging. *Hum. Brain Mapp* 41, 3235–3252. doi: 10.1002/hbm.25011

Lombardi, A., Amoroso, N., Diacono, D., Monaco, A., Logroscino, G., De Blasi, R., et al. (2020a). Association between structural connectivity and generalized cognitive spectrum in Alzheimer's disease. *Brain Sci.* 10:879. doi: 10.3390/brainsci10110879

Lombardi, A., Amoroso, N., Diacono, D., Monaco, A., Tangaro, S., and Bellotti, R. (2020b). Extensive evaluation of morphological statistical harmonization for brain age prediction. *Brain Sci.* 10:364. doi: 10.3390/brainsci10060364

Lombardi, A., Lella, E., Amoroso, N., Diacono, D., Monaco, A., Bellotti, R., et al. (2019). "Multidimensional neuroimaging processing in recas datacenter," in *International Conference on Internet and Distributed Computing Systems* (Naples: Springer), 468–477. doi: 10.1007/978-3-030-34914-1_44

Lombardi, A., Monaco, A., Donvito, G., Amoroso, N., Bellotti, R., and Tangaro, S. (2020c). Brain age prediction with morphological features using deep neural networks: results from predictive analytic competition 2019. *Front. Psychiatry* 11:1613. doi: 10.3389/fpsyt.2020.619629

Lundberg, S. M., and Lee, S.-I. (2017). "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, Vol. 30, eds I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Long Beach, CA: Curran Associates, Inc.), 4765–4774.

Magnotta, V. A., Friedman, L., and Birn, F. (2006). Measurement of signal-to-noise and contrast-to-noise in the fbirn multicenter imaging study. *J. Digit. Imaging* 19, 140–147. doi: 10.1007/s10278-006-0264-x

McGinnis, S. M., Brickhouse, M., Pascual, B., and Dickerson, B. C. (2011). Age-related changes in the thickness of cortical zones in humans. *Brain Topogr.* 24, 279–291. doi: 10.1007/s10548-011-0198-6

Meng, Y., Li, G., Lin, W., Gilmore, J. H., and Shen, D. (2014). Spatial distribution and longitudinal development of deep cortical sulcal landmarks in infants. *Neuroimage* 100, 206–218. doi: 10.1016/j.neuroimage.2014.06.004

Miller, T. (2019). Explanation in artificial intelligence: insights from the social sciences. *Artif. Intell.* 267, 1–38. doi: 10.1016/j.artint.2018.07.007

Moradi, M., and Samwald, M. (2021). *Post-hoc* explanation of black-box classifiers using confident itemsets. *Expert Syst. Appl.* 165:113941. doi: 10.1016/j.eswa.2020.113941

Peng, H., Gong, W., Beckmann, C. F., Vedaldi, A., and Smith, S. M. (2021). Accurate brain age prediction with lightweight deep neural networks. *Med. Image Anal.* 68:101871. doi: 10.1016/j.media.2020.101871

Preul, C., Hund-Georgiadis, M., Forstmann, B. U., and Lohmann, G. (2006). Characterization of cortical thickness and ventricular width in normal aging: a morphometric study at 3 Tesla. *J. Magnet. Reson. Imaging* 24, 513–519. doi: 10.1002/jmri.20665

Remer, J., Croteau-Chonka, E., Dean, D. C. III, D'Arpino, S., Dirks, H., et al. (2017). Quantifying cortical development in typically developing toddlers and young children, 1-6 years of age. *Neuroimage* 153, 246–261. doi: 10.1016/j.neuroimage.2017.04.010

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). ""Why should I trust you?": explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16* (New York, NY: Association for Computing Machinery), 1135–1144. doi: 10.1145/2939672.2939778

Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). "Anchors: high-precision model-agnostic explanations," in *Proceedings of the AAAI Conference on Artificial Intelligence* (New Orleans, LA).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice), 618–626. doi: 10.1109/ICCV.2017.74

Shaw, P., Greenstein, D., Lerch, J., Clasen, L., Lenroot, R., Gogtay, N., et al. (2006). Intellectual ability and cortical development in children and adolescents. *Nature* 440, 676–679. doi: 10.1038/nature04513

Shehzad, Z., Giavasis, S., Li, Q., Benhajali, Y., Yan, C., Yang, Z., et al. (2015). The preprocessed connectomes project quality assessment protocol-a resource for measuring the quality of MRI data. *Front. Neurosci.* 2015:47. doi: 10.3389/conf.fnins.2015.91.00047

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). "Fooling lime and shap: adversarial attacks on *post hoc* explanation methods," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY), 180–186. doi: 10.1145/3375627.3375830

Tamnes, C. K., Herting, M. M., Goddings, A.-L., Meuwese, R., Blakemore, S.-J., Dahl, R. E., et al. (2017). Development of the cerebral cortex across adolescence: a multisample study of inter-related longitudinal changes in

cortical volume, surface area, and thickness. *J. Neurosci.* 37, 3402–3412. doi: 10.1523/JNEUROSCI.3302-16.2017

Tamnes, C. K., Østby, Y., Fjell, A. M., Westlye, L. T., Due-Tønnessen, P., and Walhovd, K. B. (2010). Brain maturation in adolescence and young adulthood: regional age-related changes in cortical thickness and white matter volume and microstructure. *Cereb. Cortex* 20, 534–548. doi: 10.1093/cercor/bhp118

Van Rooij, D., Anagnostou, E., Arango, C., Auzias, G., Behrmann, M., Busatto, G. F., et al. (2018). Cortical and subcortical brain morphometry differences between patients with autism spectrum disorder and healthy individuals across the lifespan: results from the enigma asd working group. *Am. J. Psychiatry* 175, 359–369. doi: 10.1176/appi.ajp.2017.17010100

Vinke, E. J., De Groot, M., Venkatraghavan, V., Klein, S., Niessen, W. J., Ikram, M. A., et al. (2018). Trajectories of imaging markers in brain aging: the Rotterdam study. *Neurobiol. Aging* 71, 32–40. doi: 10.1016/j.neurobiolaging.2018.07.001

Vu, M.-A. T., Adalı, T., Ba, D., Buzsáki, G., Carlson, D., Heller, K., et al. (2018). A shared vision for machine learning in neuroscience. *J. Neurosci.* 38, 1601–1607. doi: 10.1523/JNEUROSCI.0508-17.2018

Wang, J., Knol, M. J., Tiulpin, A., Dubost, F., de Bruijne, M., Vernooij, M. W., et al. (2019). Gray matter age prediction as a biomarker for risk of dementia. *Proc. Natl. Acad. Sci. U.S.A.* 116, 21213–21218. doi: 10.1073/pnas.1902376116

Wang, M., Zheng, K., Yang, Y., and Wang, X. (2020). An explainable machine learning framework for intrusion detection systems. *IEEE Access* 8, 73127–73141. doi: 10.1109/ACCESS.2020.2988359

Zhao, Y., Klein, A., Castellanos, F. X., and Milham, M. P. (2019). Brain age prediction: Cortical and subcortical shape covariation in the developing human brain. *Neuroimage* 202:116149. doi: 10.1016/j.neuroimage.2019.116149

Zielinski, B. A., Prigge, M. B., Nielsen, J. A., Froehlich, A. L., Abildskov, T. J., Anderson, J. S., et al. (2014). Longitudinal changes in cortical thickness in autism and typical development. *Brain* 137, 1799–1812. doi: 10.1093/brain/awu083

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Lombardi, Diacono, Amoroso, Monaco, Tavares, Bellotti and Tangaro. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.