



Article

Explanations of Machine Learning Models in Repeated Nested Cross-Validation: An Application in Age Prediction Using Brain Complexity Features

Riccardo Sceda ¹  and Stefano Diciotti ^{1,2,*} 

¹ Department of Electrical, Electronic, and Information Engineering “Guglielmo Marconi”, University of Bologna, 47522 Cesena, Italy; riccardo.sceda2@unibo.it

² Alma Mater Research Institute for Human-Centered Artificial Intelligence, 40136 Bologna, Italy

* Correspondence: stefano.diciotti@unibo.it; Tel.: +39-0547-339121

Abstract: SHAP (Shapley additive explanations) is a framework for explainable AI that makes explanations locally and globally. In this work, we propose a general method to obtain representative SHAP values within a repeated nested cross-validation procedure and separately for the training and test sets of the different cross-validation rounds to assess the real generalization abilities of the explanations. We applied this method to predict individual age using brain complexity features extracted from MRI scans of 159 healthy subjects. In particular, we used four implementations of the fractal dimension (FD) of the cerebral cortex—a measurement of brain complexity. Representative SHAP values highlighted that the most recent implementation of the FD had the highest impact over the others and was among the top-ranking features for predicting age. SHAP rankings were not the same in the training and test sets, but the top-ranking features were consistent. In conclusion, we propose a method—and share all the source code—that allows a rigorous assessment of the SHAP explanations of a trained model in a repeated nested cross-validation setting.

Keywords: brain complexity; explainable AI; fractal dimension; machine learning; SHAP



Citation: Sceda, R.; Diciotti, S. Explanations of Machine Learning Models in Repeated Nested Cross-Validation: An Application in Age Prediction Using Brain Complexity Features. *Appl. Sci.* **2022**, *12*, 6681. <https://doi.org/10.3390/app12136681>

Academic Editor: Vincent A. Cicirello

Received: 31 May 2022

Accepted: 28 June 2022

Published: 1 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The explainability of artificial intelligence (AI) algorithms is becoming more and more critical in many fields of research [1,2], especially in medicine. In fact, clinicians have to trust algorithms when a prediction occurs and make crucial decisions that can have physical and psychological implications for patients. Therefore, an interpretable model should clarify how it arrived at a specific decision and which features it considered relevant. In medicine, interpretability is thus necessary for AI to ensure concordance with medical goals [3]. For this purpose, many explainable AI (XAI) methods have been proposed to unfold models' behavior [4–8]. For example, the *local interpretable model-agnostic explanations* (LIME) method interprets individual model predictions based on a local approximation of the model around a given prediction [7]. The *deep learning important features* (DeepLIFT) method is a recursive prediction explanation method for deep learning models [8]. It decomposes the output prediction on a specific input by backpropagating the contributions of all neurons in the network to every input feature. Then, DeepLIFT compares the activation of each neuron to its reference activation and assigns contribution scores according to this difference.

Recently, Lundberg and Lee proposed an innovative approach for XAI, called SHAP (Shapley additive explanations) [9]. SHAP is a powerful XAI framework for interpreting predictions based on classical Shapley values from game theory by assigning to each feature an importance value for every sample [10]. This approach can provide both local (on the single sample) and global explanations of the model. According to game theory, features can be seen as players playing a cooperative game to provide a specific prediction, and the importance of each feature can be computed through the so-called Shapley values [9]. The

SHAP method is preferable to other explanation methods because it satisfies three desirable properties [9]: *local accuracy* (i.e., the definition of local explanations in addition to global explanations); *missingness* (i.e., a missing feature does not contribute to the explanation); *consistency* (i.e., if a model changes so that the marginal contribution of a feature value increases or stays the same (regardless of other features), the SHAP value also increases or stays the same). A comparison between SHAP and other XAI methods (e.g., LIME) has already been explored in the literature (see, e.g., [11,12]). In particular, Lombardi et al. [11] showed that SHAP values can provide more reliable explanations (i.e., less influenced by small variations of the training set) for the morphological aging mechanisms and can be exploited to identify personalized age-related imaging biomarkers.

However, one of the main characteristics of many machine learning strategies is their inherent stochastic nature [13], which leads to the performance not being exactly reproducible. The problem of reproducibility in science has been debated for the past few decades, especially in medicine and healthcare [14–18]. For example, even the choice of random seeds in many machine learning models (whenever they are present) could lead to high variability of the performance between two different training procedures of the same model [19]. To reduce this phenomenon, many authors repeat the training procedure tens of times, changing the random seeds during the process and taking a final average performance of the model based on these repetitions (see, e.g., [20–23]). Furthermore, due to the frequent scarcity of data in medical research [24], one common approach in medicine is the repetition of a nested cross-validation (nCV) loop [25]. NCV is a procedure that helps examine the unbiased generalization performance of the trained models and, simultaneously, performs hyperparameters optimization [25]. This method is especially effective when the amount of data is relatively low because it allows for training and testing a model many times using wide non-overlapping portions (i.e., folds) of the dataset. This approach allows the possibility of testing the model on all subjects of the dataset and obtaining an average performance.

However, the SHAP framework has been proposed for hold-out strategies [26–30], where SHAP values are computed only when a final model is trained. Some authors adopted the SHAP method with CV strategies [31–36], but SHAP was used only after the CV procedure on often unclear portions of the dataset. Recently, two related works [11,37] adopted an average of SHAP values, performing multiple trainings of the model with different undersampling of the training data and computing SHAP values on the test sets. To the best of our knowledge, the application of SHAP values in a repeated nCV strategy is lacking. Since SHAP values depend on the model predictions, variability in the performance of re-trained models may lead to the variability of SHAP values, with the risk of reducing the consistency of the model's explainability. Moreover, we consider it essential to evaluate the SHAP values separately for the training and test set of the different (outer) cross-validation rounds to evaluate the generalization abilities of the SHAP explanations of a trained model.

For these reasons, in this study, we extended the use of SHAP values for a repeated nCV setting by estimating representative SHAP values separately for the training and test sets. We also aimed to share our open-source tool to enable other researchers to use SHAP explainable values in repeated nCV in their own studies. To test our method, we applied representative SHAP values computed in nCV to two regression tasks and one classification task in predicting individual age using brain complexity features from two public and international neuroimaging datasets of in vivo magnetic resonance imaging (MRI) scans for a total of 159 healthy subjects (age range 6–85 years).

2. Materials and Methods

2.1. Shapley Values

Shapley values are a concept taken from cooperative game theory and are used to attribute a player's contribution to the end result of a game [10]. Let us consider a cooperative game where a set of players each collaborate to create some value. If we can

measure the total result of the game, Shapley values capture the marginal contribution of each player to the end result. Now, let us imagine a machine learning model as a game in which features cooperate to produce a model output and associate to each feature a contribution in terms of the Shapley value. A Shapley value for a feature is computed by the difference between the prediction of the model output with that feature and the prediction of the model without that feature. This method requires re-training the model on all features subsets $S \subseteq F$, where F is the set of all features. If we consider a generic model f , we can denote a model trained with the i -th feature present as $f_{S \cup \{i\}}$ and a model trained without the i -th feature as f_S . Then, predictions from the two models are compared on the difference $f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)$, where x_S and $x_{S \cup \{i\}}$ represent the values of the input features in the set S and $S \cup \{i\}$, respectively. We need to sum this term over all the possible combinations of subsets S to get ϕ_i , the marginal value of adding the i -th feature to the training. This can be accomplished by adding the weighted average among all possible differences that give the Shapley value of the i -th feature, as follows:

$$\phi_i = \frac{1}{|F|} \sum_{S \subseteq F \setminus \{i\}} \binom{|F| - 1}{|S|}^{-1} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)], \quad (1)$$

where $|S|$ and $|F|$ are the cardinalities of S and F , respectively. The combinatorial term calculates how many permutations of each subset size we have when constructing it among all remaining features excluding feature i . We then use this combinatorial term to divide the marginal contribution of the feature i to all groups of size $|S|$. In addition, we have to divide them by the number of features participating in the model prediction, that is, the total number of features F . This term is needed to average out the effect of how much the feature i contributes regardless of the size of the total number of features. We can use these values as contributions of the features for the model output. Indeed, we can define an *explanation model* g as a linear function of the feature contributions [9]:

$$g(z) = \phi_0 + \sum_{i=1}^M \phi_i z_i, \quad (2)$$

where ϕ_0 is the SHAP value equal to $E[f(z)]$ (i.e., the average of the samples' outcomes), z_i are binary variables with $z_i \in \{0, 1\}$, M is the number of input features, and binary values refer to the presence (1) or absence (0) of a feature. Shapley values satisfy the local accuracy and consistency properties [9], but still cannot handle missing values. In order to also satisfy the missingness property, Lundberg and Lee proposed SHAP values, which are explained in Section 2.2.

2.2. SHAP Values

SHAP values are the classical Shapley values of a *conditional expectation function* of the original model f [9]. Since most models cannot handle arbitrary patterns of missing input values, Lundberg and Lee approximated missing values with values of the dataset picked randomly to cancel their statistical power [9]. For this reason, SHAP values provide a unique additive feature importance measure that adheres to all the properties, including the missingness property. A single SHAP value is a real number that refers to a single feature of a sample. The sign of the SHAP value tells us along which direction the feature drives the output of a specific sample, while the absolute value tells us the impact of that feature. The sum of the SHAP values for a given sample $\sum_{i=1}^M \phi_i = f(x) - E[f(z)]$ provides the difference between the output prediction and the base value, which is the value of a featureless model, that is, the average of the samples' outcomes $E[f(z)]$.

2.3. Computing SHAP Values in Repeated Nested Cross-Validation

Let us consider a dataset composed of N samples with M features and a K -fold nCV procedure (see Figure 1). This strategy involves nesting two K -fold CV loops, where the inner loop is used to optimize, for example, model hyperparameters. The outer loop

gives an unbiased estimate of the performance of the best model [25]. The procedure starts by splitting the dataset into K folds (outer CV): one fold is kept as a test set of the outer CV, while the other $K - 1$ folds (the training set of the outer CV) are, in turn, split into K inner folds, that is, $K - 1$ for training and the K -th for validation, to provide an unbiased evaluation of the model fit on the inner training set while tuning the model's hyperparameters. Once the best combination of hyperparameters that maximized the performance metrics in the validation set has been found, the model with that combination of hyperparameters is re-trained on the outer training set and tested on the test set kept out from the outer CV. The nested CV is repeated R times with different random seeds to make different data splitting of the K folds. This procedure can be used both for regression and classification tasks.

The pseudo-code for the computation of representative SHAP values in the training and test sets of the outer CV is illustrated in Algorithm 1. For each repetition r of the outer CV loop, we compute SHAP values ϕ_{nk}^{ir} of every sample n and feature i for the k round (split iteration) of the outer CV separately for the training and test using the SHAP Python module *Explainer* [38]. Then, for each sample n , we compute a representative SHAP value for the training $(\phi_{train})_n^{ir}$ and test $(\phi_{test})_n^{ir}$ sets. For the training set, we compute $(\phi_{train})_n^{ir}$ as the average of the SHAP values overall for the $K - 1$ folds of the outer CV as follows:

$$(\phi_{train})_n^{ir} = \frac{1}{K-1} \sum_{k=1}^{K-1} (\phi_{nk}^{ir}). \quad (3)$$

Since, in the r repeated CV, a sample n belongs to one fold used as a test only, namely fold k^* , the representative SHAP value of the test set of that sample n is simply

$$(\phi_{test})_n^{ir} = \phi_{nk^*}^{ir}. \quad (4)$$

Algorithm 1 N : number of samples; M : number of features; K : number of folds; R : number of repetitions.

```

X ← Dataset # Data table with N samples and M features (N rows × M columns)
y ← Target
model ← Classifier
Explainer ← SHAP.Explainer() # Shap function which computes SHAP values
train_folds_shap_values ← 0 # Initialized matrix (N rows × M columns)
test_folds_shap_values ← 0 # Initialized matrix (N rows × M columns)
for r in 1, ..., R do
  fold_splits ← split(K)
  innerCV(fold_splits, model)
  outerCV(innerCV, K).fit(X, y)
  for k in fold_splits do
    X_train, y_train ← fold_splits.train(k)
    X_test, y_test ← fold_splits.test(k)
    best_k_model ← outer_CV[k].best_model
    best_k_model.fit(X_train, y_train)
    train_shap_values ← Explainer(best_k_model(X_train))
    test_shap_values ← Explainer(best_k_model(X_test))
    train_folds_shap_values ← train_folds_shap_values +  $\frac{\text{train\_shap\_values}}{K-1}$ 
    test_folds_shap_values ← test_folds_shap_values + test_shap_values
  end for
end for
average_train_folds_shap_values = train_folds_shap_values / R #  $\bar{\phi}_{train}$ 
average_test_folds_shap_values = test_folds_shap_values / R #  $\bar{\phi}_{test}$ 

```

Finally, after the R repeated CVs, the final representative SHAP values for the sample n and feature i are obtained by averaging over the R repetitions in both the training and test sets:

$$(\bar{\phi}_{train})_n^i = \frac{1}{R} \sum_{r=1}^R (\phi_{train})_n^{ir}, \tag{5}$$

$$(\bar{\phi}_{test})_n^i = \frac{1}{R} \sum_{r=1}^R (\phi_{test})_n^{ir}. \tag{6}$$

2.4. Experimental Tests: Age Prediction Using Features of Brain Complexity

Individual age prediction using neuroimaging data is a popular approach for identifying biomarkers supporting brain health [39]. In this context, biomarkers quantifying brain complexity, including the local gyrification index (IGI) and the fractal dimension (FD) of the cerebral gray and white matter, have been proved to have predictive capabilities in age prediction [39–41]. In a previous paper, we used two public datasets to show that, among others, our implementation of the fractal dimension, using an automated selection of the fractal scale within which the cerebral cortex manifests the highest statistical self-similarity, yielded the most accurate machine learning models for individual age prediction [40].

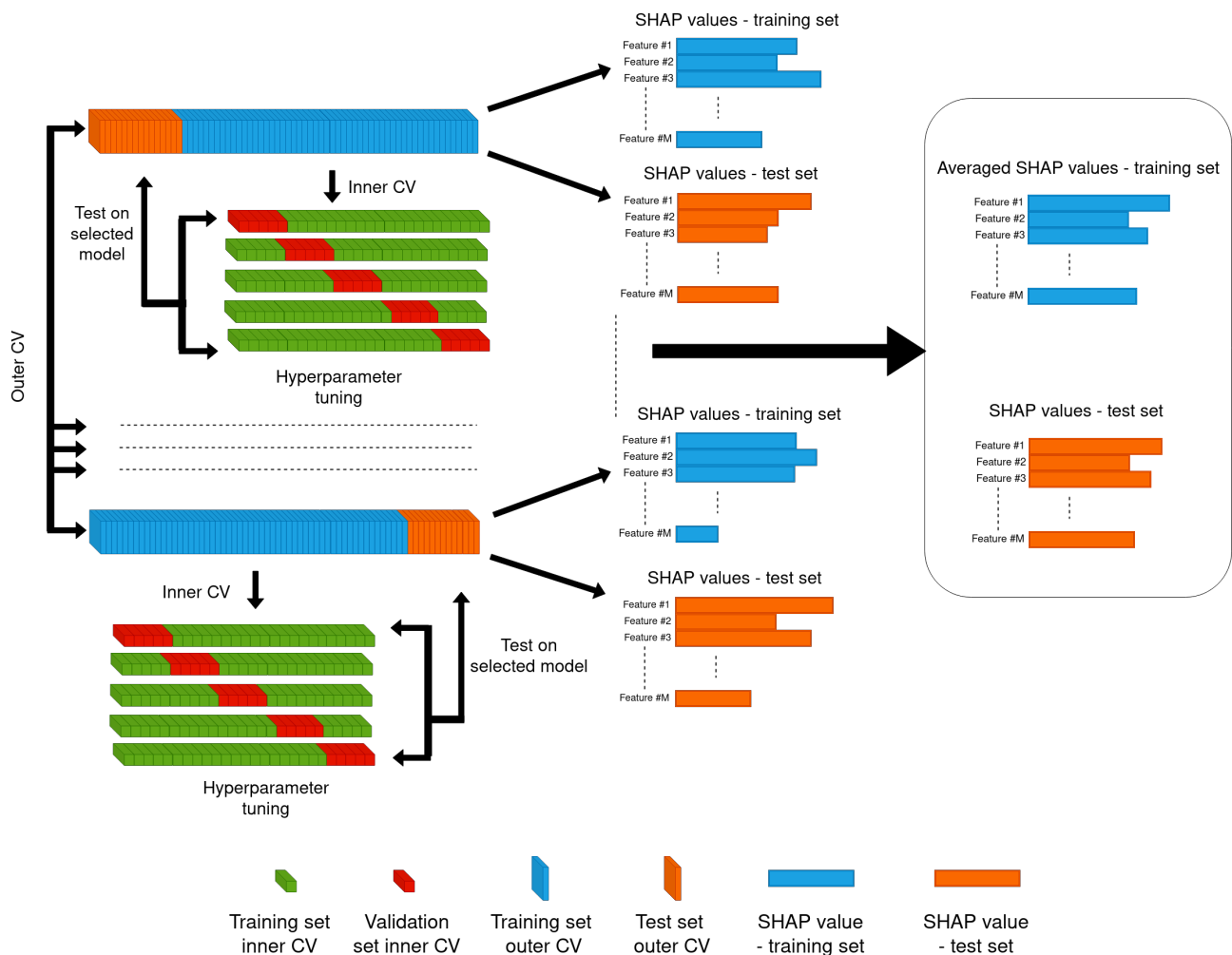


Figure 1. Schematic representation of the computation of representative SHAP values. SHAP values of training and test folds are computed separately for each round of the outer CV. Then, SHAP values are averaged over the training folds. This procedure is repeated K times, and SHAP values for the training and test sets are averaged over the R repetitions (Image adapted from [42]).

To prove the utility of computing the SHAP values in repeated nCV, in this study, we considered two regression tasks and one classification task for age prediction using features of brain complexity extracted from MRI data [40]. In particular, we used the high-resolution public and international T_1 -weighted datasets of healthy children and adolescents (Nathan Kline Institute (NKI)—Rockland Sample Pediatric Multimodal Imaging Test–Retest Sample—NKI2 dataset [43,44]) and adults (International Consortium for Brain Mapping (ICBM) dataset [45]). Briefly, the NKI2 dataset comprises MRI examinations of 73 healthy pediatric subjects aged 6 to 17 years (43 males and 30 females, age 11.8 ± 3.1 years, mean \pm standard deviation). The ICBM dataset comprises MRI examinations of 86 healthy adult and elderly subjects ranging from 19 to 85 years (41 males and 45 females, age 44.2 ± 17.1 years). For classification purposes, we also considered a dichotomous task defined as the prediction of a young group vs. an elder group in the ICBM dataset. The young group consisted of subjects with an age ≤ 30 years (25 subjects—9 males and 16 females, 22.6 ± 3.3 years), and the elder group of subjects with an age ≥ 56 years (28 subjects—11 males and 17 females, 64.9 ± 8.2 years).

The extraction of features of brain complexity from MRI data has been described in detail previously [40]. Briefly, a completely automated cortical reconstruction of each subject's structural T_1 -weighted MRI scan was performed by employing the FreeSurfer image analysis suite (<http://surfer.nmr.mgh.harvard.edu/>, accessed on 29 May 2022) [46], a dedicated brain segmentation software program [47–52]. This includes removal of non-brain tissue using a hybrid watershed/surface deformation procedure, automated Talairach transformation, segmentation of the subcortical white matter and deep gray matter volumetric structures, intensity normalization, tessellation of the gray/white matter boundary, automated topology correction [53], and surface deformation following intensity gradients to optimally place the gray/white and gray/cerebrospinal fluid borders at the location where the greatest shift in intensity defines the transition to the other tissue class. The local cortical gyrification IGI was computed following a surface-based approach [54]. Briefly, in each vertex, a spherical region of interest is delineated on an outer envelope (ROI_O) that tightly wraps the pial cortical surface, and its corresponding region of interest on the pial cortical surface (ROI_P) is identified using a matching algorithm based on geodesic constraints. Thus, the IGI is derived as the ratio between ROI_P and ROI_O areas, quantifying the amount of cortex buried within the sulcal folds in the surrounding spherical region. Then, we averaged the IGI within the entire cortex to obtain a gyrification index (GI) representative of the cortical complexity of each subject. Moreover, we recorded the following FreeSurfer outputs: the cerebral cortical gray matter volume (CortexVol), the estimated intracranial volume (eTIV), and the average cortical thickness (CT) throughout the cerebral cortex.

Lastly, we estimated the FD of the cerebral cortex of each subject using four different strategies for the selection of spatial scales. These include the use of (1) a priori selection of the interval 4 mm–256 mm (inspired by Kiselev et al. [55]) ($FD_{A\ priori\ #1}$); (2) a priori selection of 5–40% of the smallest Euclidean dimension of the cerebral cortex [56] (rounded to the nearest power of 2) ($FD_{A\ priori\ #2}$); (3) an automated selection of spatial scales, within which the cerebral cortex manifests the highest statistical self-similarity [57,58] ($FD_{Auto\ Marzi\ et\ al.\ 2018}$); (4) an improved automated selection of the interval of spatial scales, based on the search of the interval of spatial scales that presents the highest rounded R^2_{adj} coefficient, and in the case of an equal rounded R^2_{adj} coefficient, preferring the widest interval in the log–log plot ($FD_{Auto\ fractalbrain}$) [40,59,60].

We predicted individual age using an extreme gradient boosting (XGBoost) model—an XGBoost regressor or classifier for regression and classification tasks, respectively. XGBoost is a tree-based machine learning model widely used to achieve cutting-edge performance on a variety of recent machine learning challenges [61]. As inputs, we thus used nine features: the four implementations of the FD of the cerebral cortex, the volume of the cerebral cortex (i.e., CortexVol), the average cortical thickness (i.e., CT), the average gyrification index (i.e., GI), the estimated total intracranial volume (i.e., eTIV), and sex. The

models' hyperparameters were chosen from a hyperparameter space through a random search based on the average performance of the model. The hyperparameters space was defined as follows: the minimum loss reduction required to make a further partition on a leaf of the tree $\gamma \in (0.6, 0.7, 0.8)$, the subsample ratio of columns when constructing each tree $\text{colsample_bytree} \in (0.25, 0.5, 0.75, 1)$, the maximum depth of a tree $\text{max_depth} \in (2, 3, 4)$, the minimum number of instances needed to be in each node $\text{min_child_weight} \in (2, 3, 5)$, the number of decision trees $n_estimators \in (5, 10, 20, 100)$, and the ratio of training data randomly sampled prior to growing trees $\text{subsample} \in (0.1, 0.2, 0.4)$. Moreover, since SHAP has an optimized implementation for tree-based models (called *TreeExplainer*), using XGBoost, we can compute SHAP values in polynomial time, in contrast to model-agnostic explainers for this class of models [62]. We adopted a repeated (100 times) nCV strategy, and we chose a five-fold CV in both the inner and outer loops because it offers a favorable bias–variance trade-off [63].

The performance in the regression and classification tasks has been measured through the mean absolute error (MAE) and area under the receiver operating characteristic (ROC) curve (AUC), respectively. The average MAE or AUC from all repetitions was computed to get a final model assessment score.

The repository with all the source code, using Google Colab notebooks, is available on GitHub at <https://github.com/Imaging-AI-for-Health-virtual-lab/SHAP-in-repeated-nested-CV> (accessed on 29 May 2022).

3. Results

For the two regression tasks, we obtained an MAE of 1.61 ± 0.14 years (mean \pm standard deviation) in the NKI2 dataset and 12.13 ± 0.86 years in the ICBM dataset. For the classification task, we obtained an ROC AUC value of 0.881 ± 0.068 and balanced accuracy of 0.77 ± 0.06 (Figure 2). The point in the ROC curve with the minimum distance from the ideal classifier (at coordinates (0,1)) showed specificity = 0.8 and sensitivity = 0.826.

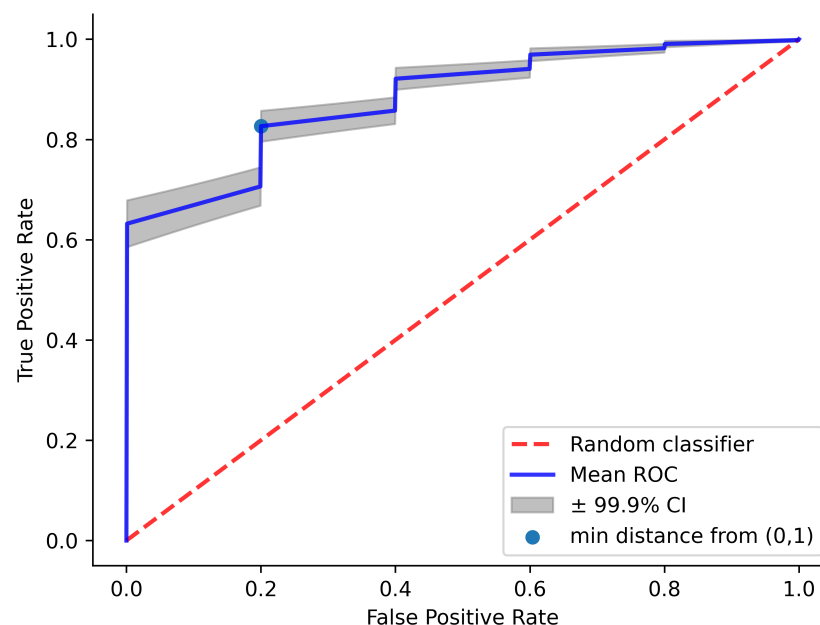


Figure 2. Average (and $\pm 99.9\%$ confidence interval (CI)) ROC curve of the model trained in the classification task in a five-fold nCV over 100 repetitions using the ICBM dataset. The point in the ROC curve with the minimum distance from the ideal classifier (at coordinates (0,1)) is represented in blue (at coordinates (0.200, 0.826)). The ROC curve of a random classifier is overlaid in red as a reference.

The beeswarm summary plots of representative SHAP values and average impact for training and test sets of the regression tasks computed using our method, in nCV over

100 repetitions, are shown in Figure 3 and 4 for the NKI2 and ICBM datasets, respectively. The beeswarm summary plots show how the top-ranking features in a dataset impact the model’s output. The given representative SHAP explanation is depicted by a single dot on each feature row for each sample (i.e., subject). The SHAP value of that feature determines the x position of the dot, and dots pile up along each feature row to show density. Color is used to display the original value of a feature [64]. The average impact is represented by bar plots showing global feature importance as the mean absolute representative SHAP value for that feature over all the given samples [65].

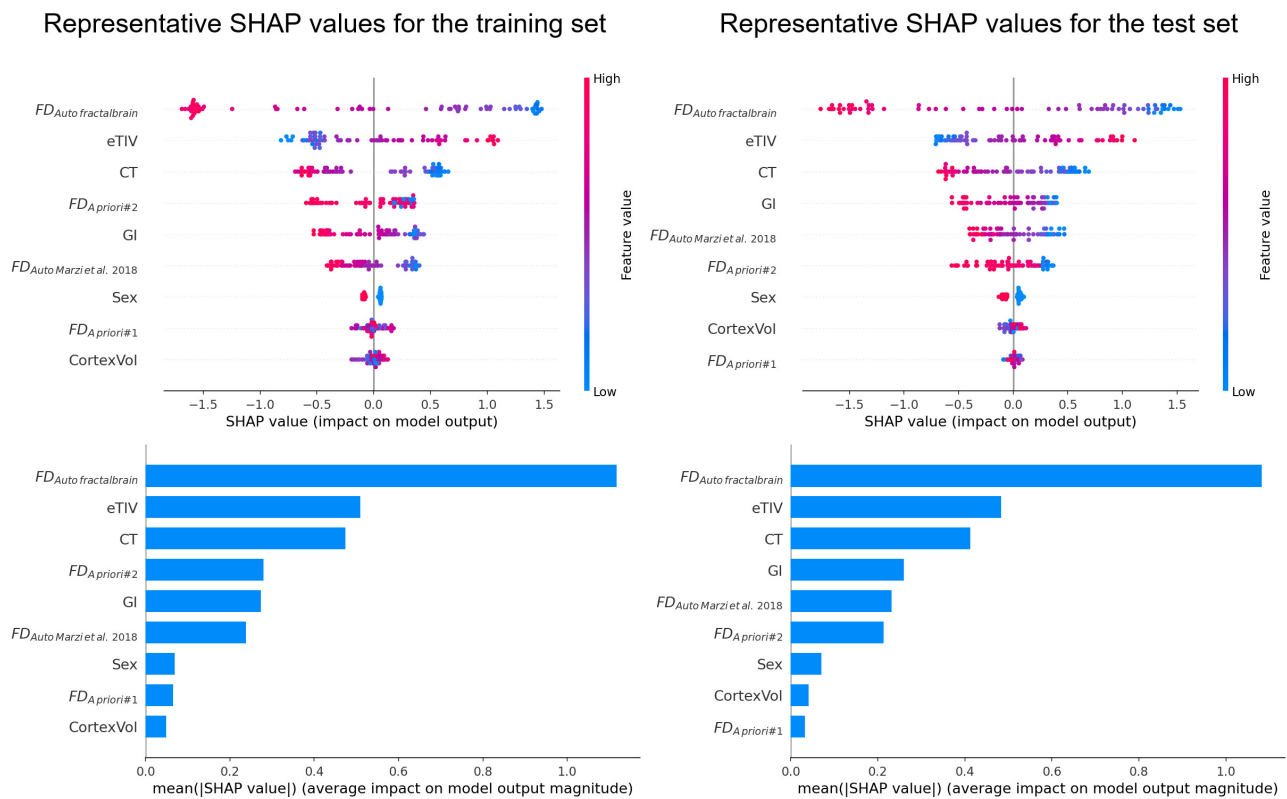


Figure 3. Results for the NKI2 regression task in a 5-fold nCV over 100 repetitions. **Top row:** beeswarm summary plots of representative SHAP values for the training (on the **left**) and test sets (on the **right**). The given SHAP explanation is represented by a single dot on each feature row for each sample (i.e., subject). The SHAP value of each feature determines the x position of the dot, and dots pile up along each feature row to show density. Color is used to display the original value of the feature. **Bottom row:** summary bar plot representing global feature importance as represented by the mean absolute SHAP value for that feature over all the given samples for the training (on the **left**) and test sets (on the **right**).

The same plots for the classification task using the ICBM dataset are shown in Figure 5.

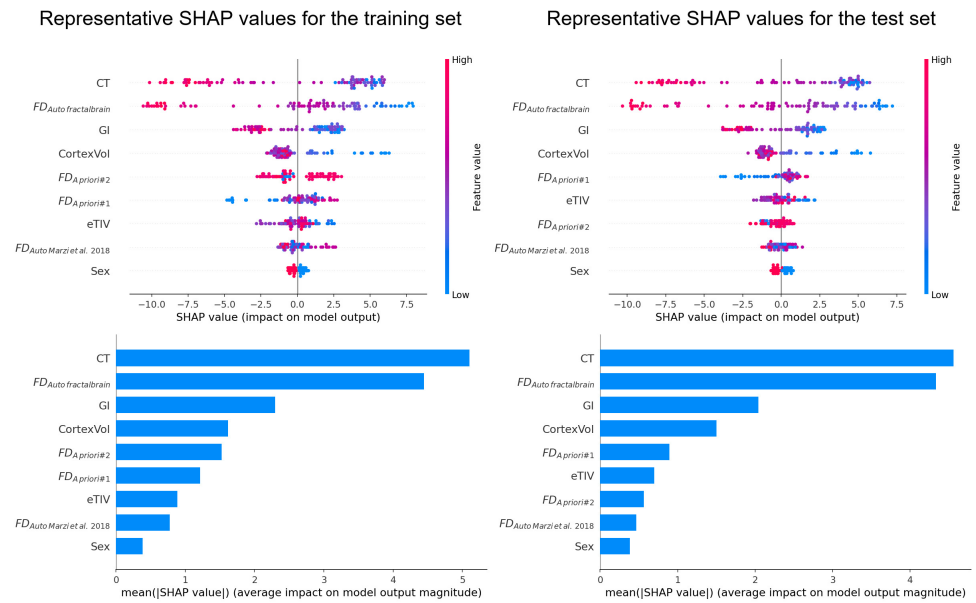


Figure 4. Results for the ICBM regression task in a five-fold nCV over 100 repetitions. **Top row:** beeswarm summary plots of representative SHAP values for the training (on the left) and test sets (on the right). The given SHAP explanation is represented by a single dot on each feature row for each sample (i.e., subject). The SHAP value of each feature determines the x position of the dot, and dots pile up along each feature row to show density. Color is used to display the original value of the feature. **Bottom row:** summary bar plot representing global feature importance as represented by the mean absolute SHAP value for that feature set over all the given samples for the training (on the left) and test sets (on the right).

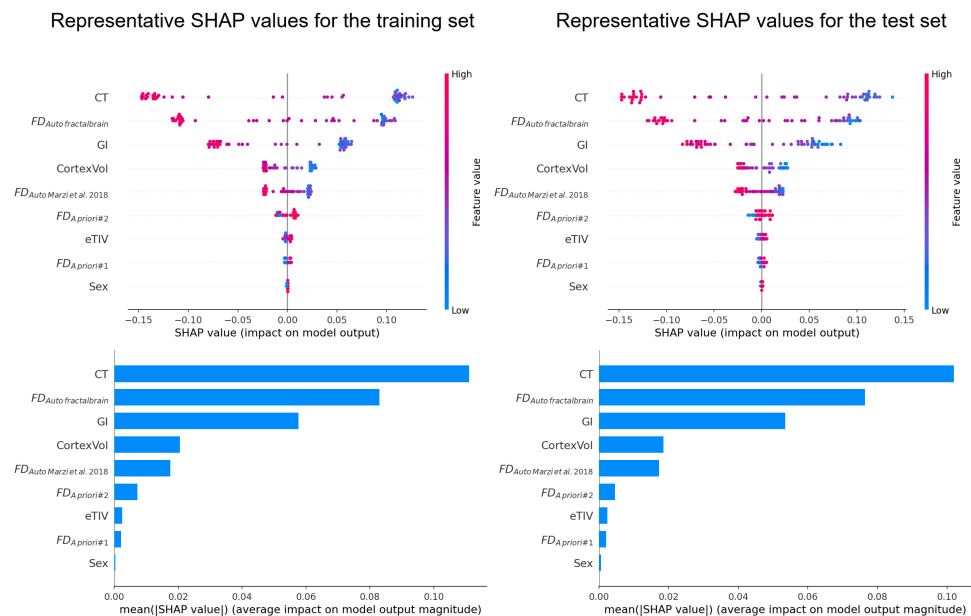


Figure 5. Results for the ICBM classification task in a five-fold nCV over 100 repetitions. **Top row:** beeswarm summary plots of representative SHAP values for the training (on the left) and test sets (on the right). The given SHAP explanation is represented by a single dot on each feature row for each sample (i.e., subject). The SHAP value of each feature determines the x position of the dot, and dots pile up along each feature row to show density. Color is used to display the original value of the feature. **Bottom row:** summary bar plot representing global feature importance as represented by the mean absolute SHAP value for that feature set over all the given samples for the training (on the left) and test sets (on the right).

4. Discussion

In this study, we proposed a method to compute representative SHAP values in a repeated nested CV procedure. Regarding the standard performance of machine learning models, in which average performance metrics are usually given, representative explainable values acting as a final assessment of the behavior of the entire model are also essential. Whereas the current literature mainly focuses on SHAP values computed on the entire dataset, we propose separate representative SHAP values for the training and test sets to allow a rigorous assessment of the generalization abilities of the SHAP explanations of a trained model.

Based on traditional Shapley values [10], SHAP uses a game-theoretic framework to reframe the task of explaining the contribution of different features to the model output for a particular instance. However, since the SHAP values depend on the model predictions, variability in the performance of re-trained models leads to variability of the model’s explainability through SHAP values. An example of this effect is shown in Figure 6, in which we report the frequency with which each feature was identified as the most important (highest impact as measured by the absolute value of the SHAP value over all the given samples) across all outer CV test folds in 100 repetitions for each regression/classification task. It is apparent that the most impactful feature changes over the different folds and repetitions. Still, the order of the most impactful features in a single iteration may differ between training and test sets—see, for example, the summary bar plot representing global feature importance for the regression task using the NKI dataset in Figure 7.

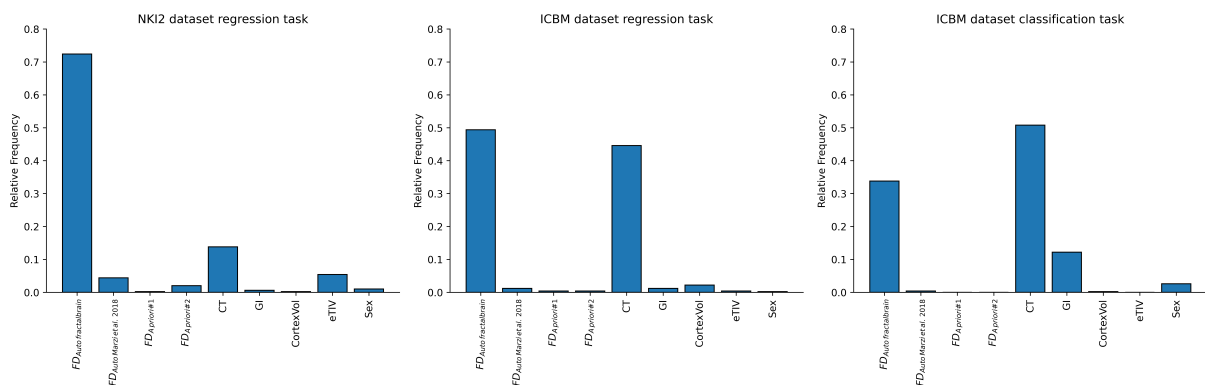


Figure 6. Relative frequency of neuroimaging features. For each regression/classification task, the frequency with which each feature was identified as the most important (highest impact as measured by the absolute value of the SHAP value over all the given samples across all outer CV test folds in 100 repetitions) is shown.

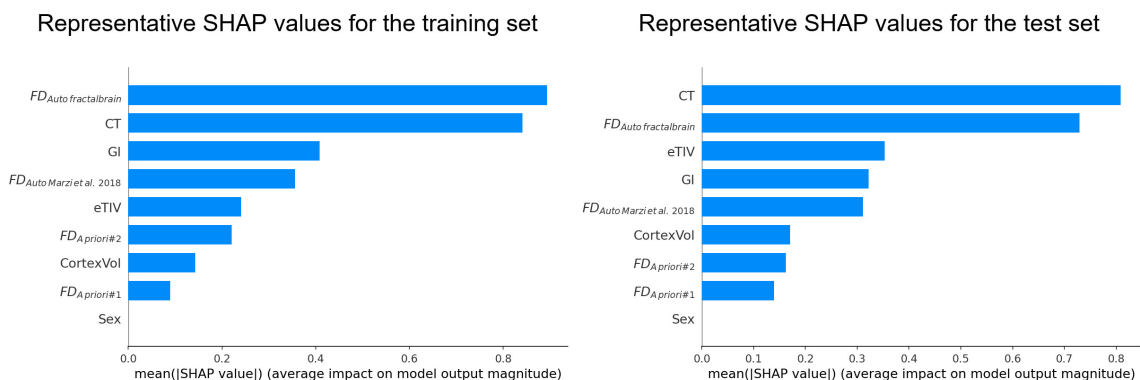


Figure 7. Summary bar plot representing global feature importance for a single iteration of the regression task using the NKI dataset, separately, for the training (on the left) and test sets (on the right). The order of the most impactful features in one iteration differs between training and test sets.

Previously, SHAP values were estimated for the test set of a single nCV repetition, thus generating potentially unstable explanations [36,66,67]. Blüthgen et al. proposed using SHAP values in the test sets of a repeated CV without detailing the procedure adopted and considering only the average impact on model output magnitude, thus losing information about the signs of the SHAP values, which inform the positive/negative association with each feature [68]. In two recent works [11,37], the average of SHAP values of samples in test sets among 100 repetitions of an ML model has been applied. In [11], the authors trained a deep neural network (DNN) model for age prediction using MRI data from the Autism Brain Imaging Data Exchange (ABIDE I) dataset collected from 17 international sites. For hyperparameter tuning, they used a leave-one-site-out CV, where the data from one site was adopted as a test set to evaluate the model's performance, while the data from all other sites was used as a training set. After each CV, they randomly undersampled the training set 100 times by removing a percentage of the samples in each iteration to produce small variations of the composition of the set and trained the DNN model to predict the subjects' age. They tested the DNN models on each test set sample, collecting 100 MAEs and SHAP values and averaging them for each sample. In [37], the authors adopted a leave-one-subject-out CV strategy using MRI data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, splitting the dataset into as many sets as the number of subjects. One subject was randomly selected for testing, while the others were used to train the model. For each CV (over 100 repetitions), they randomly undersampled the training set multiple times by selecting a fixed amount of samples for each diagnostic category from the training set. Then, a random forest model was trained within each CV round based on a grid search and nested k-fold stratified CV. The tuned model was tested on each sample of the test subject, and SHAP values were computed 100 times and averaged for each sample. Basically, in their first work, the authors trained the same model multiple times on different portions of training datasets [11], whereas, in the other study, they added hyperparameter tuning within the repetitions to select the best model given by the different subsets [37].

Our method allows the user to estimate representative SHAP values, separately for the training and test sets, in a repeated nCV setting, following a well-documented algorithm, which is accompanied by the source code. This will enable other researchers to apply the procedure in their own studies. It differs from the Lombardi et al. approach [11,37] for two main reasons: (i) we repeat the nested CV R times rather than generating repeated undersampled training set in a single nested CV; (ii) our proposed method allows the user to separately compute representative SHAP values for the training and test sets—coherently averaged, sample by sample, among folds and repetitions. This gives the user a stronger understanding of the average behavior of the model's interpretability in a very robust and popular validation setting (i.e., the repeated nested CV). Moreover, our method can be easily applied to simpler validation schemes, including repeated hold-out procedures and any machine learning model in a regression or classification task.

Our results on individual age prediction using brain complexity features are consistent with previous findings [40]. We previously showed that a monotonic decrease in structural complexity (in terms of $FD_{Auto\ fractalbrain}$) of the cerebral cortex with age during almost all the lifespan [40] and, more recently, that cardiorespiratory fitness is positively associated with cortical gray matter complexity in the temporal lobe, a region which is particularly sensitive to normal and pathological aging [60]. In the present study, as expected, low values of brain complexity ($FD_{Auto\ fractalbrain}$) and cortical thickness give a positive contribution to the model output (individual age) in both children and adults (see the beeswarm plots in Figures 3–5). In other words, we confirmed that a lower brain complexity and cortical thinning are valuable predictors of older subjects in both a young and adult cohort. Moreover, our results suggest that our latest development of the fractal dimension ($FD_{Auto\ fractalbrain}$) [40] is more predictive of individual age compared to other implementations. This result strengthens the importance of the selection of the interval of spatial scales for an adequate characterization of the structural complexity of the cerebral cortex, which is especially fascinating when using ultra-high-field MRI [59]. Moreover,

$FD_{Auto\ fractalbrain}$ was the most impactful neuroimaging feature for predicting the age in children (NKI2 dataset) and the second most important feature for adults (ICBM dataset), with an impact on the model output close to that of the first top-ranking CT feature (see Figures 4 and 5). This result confirms the ability of the fractal dimension of the cerebral cortex, in addition to cortical thickness and gyrification, to characterize brain maturation and aging, as previously observed for neurodegeneration [69]. In addition, we showed that the global feature importance of the $FD_{Auto\ fractalbrain}$ was consistently greater than that of the GI—a well-established index of the structural complexity of the human cortex.

As expected, feature rankings were not consistently the same in the training and test sets. Indeed, whereas the same ranking was observed for the ICBM regression task (see Figure 4), this was not the case for the NKI2 regression task (see Figure 3) and ICBM classification task (see Figure 5), in which only the first three and four top-ranking features were identical, respectively.

The main limitation of our proposed method is the computation time. Indeed, while the SHAP's explainer *TreeExplainer*, tailored for tree-based models such as XGBoost, is very efficient, the model-agnostic SHAP explainer is computationally demanding, especially within repeated nested cross-validation. Still, we considered the age prediction task using brain complexity features to exemplify the use of SHAP in repeated nCV. We refer to more specialized literature for improving age prediction using, for example, functional connectivity features extracted by functional MRI [70] and electroencephalography (EEG) data [71], which could also potentially use recent deep learning advances [72,73].

5. Conclusions

We proposed a method to compute representative SHAP values of the behavior of a machine learning model in a repeated nested cross-validation procedure, separately for the training and test sets. This will allow a rigorous assessment of the SHAP explanations of a trained model. Future efforts should focus on developing integrated frameworks for the training, testing, and explainability of AI models designed in machine learning pipelines, independently of the validation strategy.

Author Contributions: R.S. and S.D. developed the methodology. R.S. wrote the software and processed the data. S.D. supervised the work. R.S. and S.D. interpreted the results, and wrote and approved the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Institutional Review Board Statement: The leading institutions, at each site where the MR images were collected, received authorization from their local ethics committees.

Informed Consent Statement: The leading institutions, at each site where the MR images were collected, obtained informed consent from all participants.

Data Availability Statement: The datasets analyzed in the current study are available online: ICBM dataset: https://fcon_1000.projects.nitrc.org/fcpClassic/FcpTable.html (accessed on 29 May 2022). NKI2 dataset: https://fcon_1000.projects.nitrc.org/indi/CoRR/html/nki_2.html (accessed on 29 May 2022).

Conflicts of Interest: The authors declare that they have no competing interests.

Abbreviations

The following abbreviations are used in this manuscript:

ABIDE	Autism Brain Imaging Data Exchange
ADNI	Alzheimer's Disease Neuroimaging Initiative
AUC	Area under the curve

CortexVol	Cerebral cortical gray matter volume
CI	Confidence interval
CT	Average cortical thickness
eTIV	Estimated intracranial volume
FD	Fractal dimension
GI	Average gyrification index
ICBM	International Consortium for Brain Mapping
IGI	Local gyrification index
MAE	Mean absolute error
ML	Machine learning
MRI	Magnetic resonance imaging
nCV	Nested cross-validation
NKI	Nathan Kline Institute
ROC	Receiver operating characteristic
SHAP	Shapley additive explanations
XAI	Explainable artificial intelligence
XGBoost	Extreme gradient boosting

References

1. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **2019**, *267*, 1–38. [[CrossRef](#)]
2. Cirillo, D.; Catuara-Solarz, S.; Morey, C.; Guney, E.; Subirats, L.; Mellino, S.; Gigante, A.; Valencia, A.; Rementeria, M.J.; Chadha, A.S.; et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ Digit. Med.* **2020**, *3*, 1–11. [[CrossRef](#)]
3. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [[CrossRef](#)]
4. Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.R.; Samek, W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS ONE* **2015**, *10*, e0130140. [[CrossRef](#)]
5. Lipovetsky, S.; Conklin, M. Analysis of regression in game theory approach. *Appl. Stoch. Model. Bus. Ind.* **2001**, *17*, 319–330. [[CrossRef](#)]
6. Štrumbelj, E.; Kononenko, I. Explaining Prediction Models and Individual Predictions with Feature Contributions. *Knowl. Inf. Syst.* **2014**, *41*, 647–665. [[CrossRef](#)]
7. Ribeiro, M.T.; Singh, S.; Guestrin, C. “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), San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 1135–1144. [[CrossRef](#)]
8. Shrikumar, A.; Greenside, P.; Kundaje, A. Learning Important Features Through Propagating Activation Differences. *arXiv* **2019**, arXiv:1704.02685.
9. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. *arXiv* **2017**, arXiv:1705.07874.
10. Shapley, L.S. 17. A Value for n-Person Games. In *Contributions to the Theory of Games (AM-28)*; Kuhn, H.W., Tucker, A.W., Eds.; Princeton University Press: Princeton, NJ, USA, 2016; Volume 2, pp. 307–318. [[CrossRef](#)]
11. Lombardi, A.; Diacono, D.; Amoroso, N.; Monaco, A.; Tavares, J.M.R.S.; Bellotti, R.; Tangaro, S. Explainable Deep Learning for Personalized Age Prediction with Brain Morphology. *Front. Neurosci.* **2021**, *15*, 674055. [[CrossRef](#)]
12. Antwarg, L.; Miller, R.M.; Shapira, B.; Rokach, L. Explaining anomalies detected by autoencoders using Shapley Additive Explanations. *Expert Syst. Appl.* **2021**, *186*, 115736. [[CrossRef](#)]
13. Sabuncu, M.R. Intelligence plays dice: Stochasticity is essential for machine learning. *arXiv* **2020**, arXiv:2008.07496.
14. Beam, A.L.; Manrai, A.K.; Ghassemi, M. Challenges to the Reproducibility of Machine Learning Models in Health Care. *JAMA* **2020**, *323*, 305–306. [[CrossRef](#)]
15. Rajpurkar, P.; Chen, E.; Oishi, B.; Banerjee, O.; Topol, E.J. AI in health and medicine. *Nat. Med.* **2022**, *28*, 31–38. [[CrossRef](#)]
16. Haibe-Kains, B.; Adam, G.A.; Hosny, A.; Khodakarami, F. Matters arising Transparency and reproducibility in artificial intelligence. *Nature* **2020**, *586*, E14–E16. [[CrossRef](#)]
17. Stower, H. Transparency in medical AI. *Nat. Med.* **2020**, *26*, 14–16. [[CrossRef](#)]
18. Walsh, I.; Fishman, D.; Garcia-Gasulla, D.; Titma, T.; Pollastri, G.; Capriotti, E.; Casadio, R.; Capella-Gutierrez, S.; Cirillo, D.; Conte, A.D.; et al. DOME: Recommendations for supervised machine learning validation in biology. *Nat. Methods* **2021**, *18*, 1122–1127. [[CrossRef](#)]
19. Amir, S.; van de Meent, J.; Wallace, B.C. On the Impact of Random Seeds on the Fairness of Clinical Classifiers. *arXiv* **2021**, arXiv:2104.06338.
20. Li, X.; Yin, B.; Tian, W.; Sun, Y. Performance of Repeated Cross Validation for Machine Learning Models in Building Energy Analysis. In Proceedings of the 11th International Symposium on Heating, Ventilation and Air Conditioning (ISHVAC 2019), Harbin, China, 12–15 July 2019; Wang, Z., Zhu, Y., Wang, F., Wang, P., Shen, C., Liu, J., Eds.; Springer: Singapore, 2020; pp. 523–531. [[CrossRef](#)]

21. Kim, J.H. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Comput. Stat. Data Anal.* **2009**, *53*, 3735–3745. [CrossRef]
22. Burman, P. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika* **1989**, *76*, 503–514. [CrossRef]
23. Vanwinckelen, G.; Blockeel, H. On Estimating Model Accuracy with Repeated Cross-Validation. 2012. Available online: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKewjc5dre_pf1AhUtNOwKHUpQClcQFn_oECBEQAQ&url=https%3A%2F%2Fflirias.kuleuven.be%2Fretrieve%2F186558%2F&usg=AOvVaw3sAhjDtQ0B2NwGcalWuwpk (accessed on 29 May 2022).
24. Lee Choong Ho, Y.H.J. Medical big data: Promise and challenges. *Kidney Res. Clin. Pract.* **2017**, *36*, 3–11. [CrossRef]
25. Mueller, A.; Guido, S. *Introduction to machine Learning with Python: A guide for Data Scientists*; O'Reilly Media: Newton, MA, USA, 2017.
26. Batunacun; Wieland, R.; Lakes, T.; Nendel, C. Using Shapley additive explanations to interpret extreme gradient boosting predictions of grassland degradation in Xilingol, China. *Geosci. Model Dev.* **2021**, *14*, 1493–1510. [CrossRef]
27. Bi, Y.; Xiang, D.; Ge, Z.; Li, F.; Jia, C.; Song, J. Prediction of the development of acute kidney injury following cardiac surgery by machine learning. *Crit. Care* **2020**, *24*, 478. [CrossRef]
28. Kim, Y.; Kim, Y. Explainable heat-related mortality with random forest and SHapley Additive exPlanations (SHAP) models. *Sustain. Cities Soc.* **2022**, *79*, 103677. [CrossRef]
29. Chen, T.; Xu, J.; Ying, H.; Chen, X.; Feng, R.; Fang, X.; Gao, H.; Wu, J. Prediction of Extubation Failure for Intensive Care Unit Patients Using Light Gradient Boosting Machine. *IEEE Access* **2019**, *7*, 150960–150968. [CrossRef]
30. Rodríguez-Pérez, R.; Bajorath, J. Interpretation of Compound Activity Predictions from Complex Machine Learning Models Using Local Approximations and Shapley Values. *J. Med. Chem.* **2020**, *63*, 8761–8777. [CrossRef]
31. Parsa, A.B.; Movahedi, A.; Taghipour, H.; Derrible, S.; Mohammadian, A.K. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accid. Anal. Prev.* **2020**, *136*, 105405. [CrossRef]
32. Bi, Y.; Xiang, D.; Ge, Z.; Li, F.; Jia, C.; Song, J. An Interpretable Prediction Model for Identifying N7-Methylguanosine Sites Based on XGBoost and SHAP. *Mol. Ther.-Nucleic Acids* **2020**, *22*, 362–372. [CrossRef]
33. Feng, D.C.; Wang, W.J.; Mangalathu, S.; Taciroglu, E. Interpretable XGBoost-SHAP Machine-Learning Model for Shear Strength Prediction of Squat RC Walls. *J. Struct. Eng.* **2021**, *147*, 04021173. [CrossRef]
34. Deb, D.; Smith, R.M. Application of Random Forest and SHAP Tree Explainer in Exploring Spatial (In)Justice to Aid Urban Planning. *ISPRS Int. J.-Geo-Inf.* **2021**, *10*, 629. [CrossRef]
35. Wang, K.; Tian, J.; Zheng, C.; Yang, H.; Ren, J.; Liu, Y.; Han, Q.; Zhang, Y. Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and SHAP. *Comput. Biol. Med.* **2021**, *137*, 104813. [CrossRef]
36. El-Sappagh, S.; Alonso, J.M.; Islam, S.; Sultan, A.M.; Kwak, K.S. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. *Sci. Rep.* **2021**, *11*, 1–26. [CrossRef]
37. Lombardi, A.; Diacono, D.; Amoroso, N.; Biecek, P.; Monaco, A.; Bellantuono, L.; Pantaleo, E.; Logroscino, G.; Blasi, R.; Tangaro, S.; et al. A robust framework to investigate the reliability and stability of explainable artificial intelligence markers of Mild Cognitive Impairment and Alzheimer's Disease. *Res. Sq.* **2022**. [CrossRef]
38. Lundberg, S.M. SHAP Explainer. 2018. Available online: https://shap.readthedocs.io/en/latest/example_notebooks/tabular_examples/tree_based_models/Python%20Version%20of%20Tree%20SHAP.html#Python-TreeExplainer (accessed on 29 May 2022).
39. Franke, K.; Gaser, C. Ten Years of BrainAGE as a Neuroimaging Biomarker of Brain Aging: What Insights Have We Gained? *Front. Neurol.* **2019**, *10*, 789. [CrossRef]
40. Marzi, C.; Giannelli, M.; Tessa, C.; Mascalchi, M.; Diciotti, S. Toward a more reliable characterization of fractal properties of the cerebral cortex of healthy subjects during the lifespan. *Sci. Rep.* **2020**, *10*, 16957. [CrossRef]
41. Madan, C.R.; Kensinger, E.A. Cortical complexity as a measure of age-related brain atrophy. *NeuroImage* **2016**, *134*, 617–629. [CrossRef]
42. Yagis, E.; Atnafu, S.W.; García Seco de Herrera, A.; Marzi, C.; Scheda, R.; Giannelli, M.; Tessa, C.; Citi, L.; Diciotti, S. Effect of data leakage in brain MRI classification using 2D convolutional neural networks. *Sci. Rep.* **2021**, *11*, 1–13. [CrossRef]
43. Nooner, K.B.; Colcombe, S.J.; Tobe, R.H.; Mennes, M.; Benedict, M.M.; Moreno, A.L.; Panek, L.J.; Brown, S.; Zavitz, S.T.; Li, Q.; et al. The NKI-Rockland Sample: A Model for Accelerating the Pace of Discovery Science in Psychiatry. *Front. Neurosci.* **2012**, *6*, 152. [CrossRef]
44. Zuo, X.N.; Anderson, J.S.; Bellec, P.; Birn, R.M.; Biswal, B.B.; Blautzik, J.; Breitner, J.; Buckner, R.L.; Calhoun, V.D.; Castellanos, F.X.; et al. An open science resource for establishing reliability and reproducibility in functional connectomics. *Sci. Data* **2014**, *1*, 1–13. [CrossRef]
45. Kötter, R.; Mazziotta, J.; Toga, A.; Evans, A.; Fox, P.; Lancaster, J.; Zilles, K.; Woods, R.; Paus, T.; Simpson, G.; et al. A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). *Philos. Trans. R. Soc. Lond. Ser. Biol. Sci.* **2001**, *356*, 1293–1322. [CrossRef]
46. Fischl, B. FreeSurfer. *NeuroImage* **2012**, *62*, 774–781. [CrossRef]

47. Rosas, H.; Liu, A.; Hersch, S.; Glessner, M.; Ferrante, R.; Salat, D.; van Der Kouwe, A.; Jenkins, B.; Dale, A.; Fischl, B. Regional and progressive thinning of the cortical ribbon in Huntington's disease. *Neurology* **2002**, *58*, 695–701. [[CrossRef](#)]
48. Han, X.; Jovicich, J.; Salat, D.; van der Kouwe, A.; Quinn, B.; Czanner, S.; Busa, E.; Pacheco, J.; Albert, M.; Killiany, R.; et al. Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. *NeuroImage* **2006**, *32*, 180–194. [[CrossRef](#)]
49. Lee, J.K.; Lee, J.M.; Kim, J.S.; Kim, I.Y.; Evans, A.C.; Kim, S.I. A novel quantitative cross-validation of different cortical surface reconstruction algorithms using MRI phantom. *NeuroImage* **2006**, *31*, 572–584. [[CrossRef](#)]
50. Kang, X.; Herron, T.J.; Cate, A.D.; Yund, E.W.; Woods, D.L. Hemispherically-Unified Surface Maps of Human Cerebral Cortex: Reliability and Hemispheric Asymmetries. *PLoS ONE* **2012**, *7*, 1–15. [[CrossRef](#)]
51. Keller, S.S.; Ahrens, T.; Mohammadi, S.; Gerdes, J.S.; Möddel, G.; Kellinghaus, C.; Kugel, H.; Weber, B.; Ringelstein, E.B.; Deppe, M. Voxel-Based Statistical Analysis of Fractional Anisotropy and Mean Diffusivity in Patients with Unilateral Temporal Lobe Epilepsy of Unknown Cause. *J. Neuroimaging* **2013**, *23*, 352–359. [[CrossRef](#)]
52. King, R.D. Computation of local fractal dimension values of the human cerebral cortex. *Appl. Math.* **2014**, *2014*, 1733–1740. [[CrossRef](#)]
53. Fischl, B.; Liu, A.; Dale, A. Automated manifold surgery: Constructing geometrically accurate and topologically correct models of the human cerebral cortex. *IEEE Trans. Med. Imaging* **2001**, *20*, 70–80. [[CrossRef](#)]
54. Schaer, M.; Cuadra, M.B.; Tamarit, L.; Lazezras, F.; Eliez, S.; Thiran, J.P. A Surface-Based Approach to Quantify Local Cortical Gyrfication. *IEEE Trans. Med. Imaging* **2008**, *27*, 161–170. [[CrossRef](#)]
55. Kiselev, V.G.; Hahn, K.R.; Auer, D.P. Is the brain cortex a fractal? *Neuroimage* **2003**, *20*, 1765–1774. [[CrossRef](#)]
56. Goñi, J.; Sporns, O.; Cheng, H.; Aznárez-Sanado, M.; Wang, Y.; Josa, S.; Arrondo, G.; Mathews, V.P.; Hummer, T.A.; Kronenberger, W.G.; et al. Robust estimation of fractal measures for characterizing the structural complexity of the human brain: Optimization and reproducibility. *Neuroimage* **2013**, *83*, 646–657. [[CrossRef](#)]
57. Marzi, C.; Ciulli, S.; Giannelli, M.; Ginestroni, A.; Tessa, C.; Mascalchi, M.; Diciotti, S. Structural complexity of the cerebellum and cerebral cortex is reduced in spinocerebellar ataxia type 2. *J. Neuroimaging* **2018**, *28*, 688–693. [[CrossRef](#)]
58. Pantoni, L.; Marzi, C.; Poggesi, A.; Giorgio, A.; De Stefano, N.; Mascalchi, M.; Inzitari, D.; Salvadori, E.; Diciotti, S. Fractal dimension of cerebral white matter: A consistent feature for prediction of the cognitive performance in patients with small vessel disease and mild cognitive impairment. *Neuroimage Clin.* **2019**, *24*, 101990. [[CrossRef](#)]
59. Marzi, C.; Giannelli, M.; Tessa, C.; Mascalchi, M.; Diciotti, S. Fractal Analysis of MRI Data at 7 T: How Much Complex Is the Cerebral Cortex? *IEEE Access* **2021**, *9*, 69226–69234. [[CrossRef](#)]
60. Pani, J.; Marzi, C.; Stensvold, D.; Wisløff, U.; Håberg, A.K.; Diciotti, S. Longitudinal study of the effect of a 5-year exercise intervention on structural brain complexity in older adults. A Generation 100 substudy. *NeuroImage* **2022**, *256*, 119226. [[CrossRef](#)]
61. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16), San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794. [[CrossRef](#)]
62. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 2522–5839. [[CrossRef](#)]
63. Hastie, T.; Tibshirani, R.; Friedman, J.H. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*; Springer: Berlin/Heidelberg, Germany, 2013.
64. Lundberg, S.M. SHAP Beeswarm Plot. 2018. Available online: https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/plots/beeswarm.html#A-simple-beeswarm-summary-plot (accessed on 29 May 2022).
65. Lundberg, S.M. SHAP Bar Plot. 2018. Available online: https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/plots/bar.html (accessed on 29 May 2022).
66. Beebe-Wang, N.; Okeson, A.; Althoff, T.; Lee, S.I. Efficient and Explainable Risk Assessments for Imminent Dementia in an Aging Cohort Study. *IEEE J. Biomed. Health Inf.* **2021**, *25*, 2409–2420. [[CrossRef](#)]
67. Siciarz, P.; Alfaifi, S.; Uytven, E.V.; Rathod, S.; Koul, R.; McCurdy, B. Machine learning for dose-volume histogram based clinical decision-making support system in radiation therapy plans for brain tumors. *Clin. Transl. Radiat. Oncol.* **2021**, *31*, 50–57. [[CrossRef](#)]
68. Blüthgen, C.; Patella, M.; Euler, A.; Baessler, B.; Martini, K.; von Spiczak, J.; Schneider, D.; Opitz, I.; Frauenfelder, T. Computed tomography radiomics for the prediction of thymic epithelial tumor histology, TNM stage and myasthenia gravis. *PLoS ONE* **2021**, *16*, 1–16. [[CrossRef](#)]
69. King, R.D.; Brown, B.; Hwang, M.; Jeon, T.; George, A.T.; Initiative, A.D.N. Fractal dimension analysis of the cortical ribbon in mild Alzheimer's disease. *Neuroimage* **2010**, *53*, 471–479. [[CrossRef](#)]
70. Monti, R.P.; Gibberd, A.; Roy, S.; Nunes, M.; Lorenz, R.; Leech, R.; Ogawa, T.; Kawanabe, M.; Hyvärinen, A. Interpretable brain age prediction using linear latent variable models of functional connectivity. *PLoS ONE* **2020**, *15*, e0232296. [[CrossRef](#)]
71. Al Zoubi, O.; Ki Wong, C.; Kuplicki, R.T.; Yeh, H.w.; Mayeli, A.; Refai, H.; Paulus, M.; Bodurka, J. Predicting Age From Brain EEG Signals—A Machine Learning Approach. *Front. Aging Neurosci.* **2018**, *10*. [[CrossRef](#)]

-
72. Zhang, X.; Yao, L.; Wang, X.; Monaghan, J.; McAlpine, D.; Zhang, Y. A survey on deep learning-based non-invasive brain signals: recent advances and new frontiers. *J. Neural Eng.* **2021**, *18*, 031002. [[CrossRef](#)]
 73. Zhao, K.; Duka, B.; Xie, H.; Oathes, D.J.; Calhoun, V.; Zhang, Y. A dynamic graph convolutional neural network framework reveals new insights into connectome dysfunctions in ADHD. *NeuroImage* **2022**, *246*, 118774. [[CrossRef](#)]