ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

INGEGNERIA BIOMEDICA, ELETTRICA E DEI SISTEMI

Ciclo 36

Settore Concorsuale: 09/G2 - BIOINGEGNERIA

Settore Scientifico Disciplinare: ING-INF/06 - BIOINGEGNERIA ELETTRONICA E INFORMATICA

DEVELOPMENT OF EXPLAINABLE AND REPRODUCIBLE ARTIFICIAL INTELLIGENCE FOR MEDICINE

Presentata da: Riccardo Scheda

Coordinatore Dottorato

Michele Monaci

Supervisore

Stefano Diciotti

Co-supervisore

Mauro Ursino

Esame finale anno 2024

Abstract

Artificial intelligence (AI) holds the potential to revolutionize medicine and healthcare, especially in diagnosis and treatment. However, integrating AI into medicine presents several challenges that demand immediate consideration. This study examines three key aspects: explainability, reproducibility, and the scarcity of data due to privacy concerns. Explainability is vital for increasing trust in AI systems, especially in medical applications where decisions directly impact patient well-being. Reproducibility ensures the reliability of machine learning models across different settings. In this work, a new algorithm is proposed to compute average explanations to enhance these aspects. This approach aims to provide consistent and reproducible explanations, particularly in validation settings, contributing to the transparency and reliability of AI in medical decision-making. Additionally, privacy regulations intensify the scarcity of medical data, which prevents the development of effective AI models. In response, this investigation explores the potential of applying swarm learning (SL). Swarm learning is a recently proposed technology that empowers collaborative model training across decentralized data and computational sources while preserving data privacy. This innovative approach overcomes data scarcity issues and ensures compliance with stringent privacy regulations, preparing for a more robust AI development in the medical domain. This study underscores the necessity of addressing critical aspects such as explainability, reproducibility, and privacy concerns when deploying AI for healthcare applications.

Contents

Abbreviations

The following abbreviations are used in this manuscript:

ABIDE	Autism Brain Imaging Data Exchange
ADOS	Autism Diagnostic Observation Schedule
ADNI	Alzheimer's Disease Neuroimaging Initiative
AI	artificial intelligence
AUC	area under the curve
CortexVol	cerebral cortical gray matter volume
CI	confidence interval
CT	average cortical thickness
CV	cross-validation
EHR	electronic health record
EPVS	enlarged perivascular spaces
eTIV	estimated intracranial volume
FA	fractional anisotropy
FD	fractal dimension
FL	federated learning
GI	average gyrification index
GM	gray matter
ICBM	international consortium for brain mapping
lGI	local gyrification index
LIME	local interpretable model-agnostic explanations
MAE	mean absolute error
MD	mean diffusivity
ML	machine learning
MRI	magnetic resonance imaging
nCV	nested cross-validation
NKI	Nathan Kline institute
ReLU	rectified linear unit
ROC	receiver operating characteristic
ROI	region of interest
SD	standard deviation
SHAP	Shapley additive explanations
SL	swarm learning
SVD	small vessel disease
TD	typical development
URI	uniform resource identifier
WM	white matter
XAI	explainable artificial intelligence
XGBoost	extreme gradient boosting

Chapter 1

Introduction

1.1 Critical aspects of artificial intelligence in medicine

In recent years, the spread of artificial intelligence has increased in many research fields. Applications of machine learning (ML) methods have been widely used to solve various complex challenges across multiple application areas, such as medical, financial, environmental, marketing, security, and industrial applications [Shehab et al., 2022]. In the healthcare context, the role of AI is becoming crucial year after year: more computationally efficient algorithms now offer unique opportunities to enhance diagnosis and improve approaches to precision medicine. Despite the enormous potential shown by AI in research, its deployment in the real world still needs to be improved. Indeed, with the increase in the use of artificial intelligence, significant challenges are arising [Saw and Ng, 2022]. These critical aspects can be summarized in three key points:

- *Explainability*: AI models are often difficult to trust and understand;
- *Reproducibility*: the results of AI models proposed in published paper journals are difficult to reproduce;
- *Privacy and Scarcity of Data*: Data are often scarce due to the rareness of diseases or privacy issues.

Indeed, in recent years, a growing amount of scientific literature has claimed a need for reproducible and explainable artificial intelligence in medicine [Loftus et al., 2022, Ciobanu-Caraus et al., 2024, Moassefi et al., 2023].

The role of explainability is crucial in healthcare and related fields because both patients and clinicians need to have tools to trust machine learning models.

The reproducibility of machine learning models is crucial to reproduce models many times, and we need to have the same results for the same experiments and data.

The scarcity of medical data is often due to privacy issues but also due to the rareness of the diseases, which leads to having too low of an amount of data eligible for the training of robust machine learning models.

In this work, we focused our research activities on these critical aspects of artificial intelligence in medicine.

Explainability

Explainability is the branch of artificial intelligence that focuses on making machine learning models clear and trustworthy. Indeed, ML models are often considered to be *black box models*, in which it is difficult or impossible to argue why a model is making a specific decision or had that particular prediction. For this purpose, many techniques of explaining models were born in the past years. Here we present three of the most known techniques used for tabular data in healthcare [Di Martino and Delmastro, 2023]:

- Deep Learning Important FeaTures (DeepLIFT) [Shrikumar et al., 2017]: a recursive prediction explanation method for deep learning models.
- Local Interpretable Model-agnostic Explanations (LIME) [Ribeiro et al., 2016]: interprets individual model predictions based on a local approximation of the model around a given prediction.
- SHapley Additive eXplanations (SHAP) [Lundberg and Lee, 2017a]: a framework for interpreting predictions based on classical Shapley values from game theory by assigning to each feature an importance value for every sample.

In the next chapters, we will introduce these three methods for enhancing the interpretability of machine learning models. Notably, our exploration will focus on adopting SHAP as a preferred method for rendering machine learning models transparent and interpretable. We will focus on the unique attributes of SHAP that position it as a powerful tool in unraveling the decision-making processes of complex models, emphasizing its capacity to offer insightful and intuitive explanations.

Reproducibility

One of the challenges in machine learning research is to ensure that presented and published results are consistent and reliable [Pineau et al., 2021]. Reproducibility, that is, obtaining similar results as shown in a scientific paper using the same code and data (when available), is necessary to verify the reliability of research findings. The training of many machine learning models makes use of *randomness*, and this is especially true for deep learning models. One attempt at creating trustworthy analyses with machine-learning models revolves around reporting analysis details such as hyperparameter values, model architectures, and data-splitting procedures. Unfortunately, such reporting requirements are insufficient to make analyses trustworthy.

For machine learning models in the life sciences to become trusted, scientists must prioritize computational reproducibility [Heil et al., 2021]. To improve 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 [Li et al., 2020, Kim, 2009, Burman, 1989, Vanwinckelen and Blockeel, 2012]. Moreover, when the datasets available have a low number of samples (which is very common in medical research), it is recommended to use validation methods to get the maximum possible amount of information using all the samples during the training process. One approach in Medicine is the repetition of a nested cross-validation (nCV) loop [Mueller and Guido, 2017]. NCV is a procedure that helps examine the unbiased generalization performance of the trained models and simultaneously performs hyperparameters optimization [Mueller and Guido, 2017]. This class of methodologies allows authors to assess more robust, consistent, and reproducible models.

Data privacy issues

Recent medical applications are primarily dominated by ML models to assist expert decisions, leading to innovations in radiology, genomics, and modern healthcare systems in general [Aouedi et al., 2023]. Despite the profitable usage of AI-based algorithms, these data-driven methods often need help with issues such as the scarcity and privacy of user data and the difficulty of institutions exchanging medical information. Several machine learning and deep learning algorithms have been applied to facilitate clinical diagnosis, but such tools often require large clinical datasets for training [Chishti et al., 2020, Al'Aref et al., 2018]. In medicine and healthcare,

one of the significant problems for artificial intelligence is the need for large amounts of data [Saw and Ng, 2022]. Limited samples often characterize single-center studies in the medical domain due to the complexity and high costs of patient data collection [Shaikhina and Khovanova, 2017]. On the one hand, this is due to the rareness of diseases [Holzinger, 2018], leading to the hospitals having a meager amount of patient samples to analyze. On the other hand, scarcity of data is given by the fact that hospitals do not want or are not allowed to share their data due to privacy policies [Rieke et al., 2020]. This leads hospitals to train machine learning models on very restricted datasets. In conventional centralized approaches, aggregating diverse datasets for comprehensive learning is impeded by strict privacy protocols, making it challenging to develop robust and generalizable models. To address this issues, in the recent years data augmentation techniques and generative AI in general has become more and more popular in the medical field [Chlap et al., 2021], giving clinicians the possibility to train models on big synthetic datasets. However, data augmentation techniques may be limited by serious challenges in building an appropriate generative model given the high intercorrelation of the medical data [Murtaza et al., 2023].

An other approach to face the scarcity of data is the use of decentralized learning, where models in different locations are trained on their local data and merged to obtain a final model. Recently, *swarm learning* appeared as a valuable approach in the medical field, allowing patient data to stay where it is generated. Swarm learning is a decentralized and collaborative approach to machine learning where a group of individual machine learning models, called *nodes* work together to collectively solve a problem or perform a task without centralized control [Warnat-Herresthal et al., 2021]. Unlike traditional machine learning models that rely on a central server or coordinator, swarm learning distributes the learning process across multiple nodes. Each node independently processes data and contributes to the overall learning task. Nodes in a SL system communicate and share information. This collaboration allows the swarm to benefit from individual nodes' different sample information, leading to improved overall performance. Swarm learning addresses privacy concerns by allowing data to remain on individual nodes: instead of sharing raw data, nodes exchange model updates or aggregated information, enabling collaborative learning without compromising the privacy of individual data. Swarm learning is inherently scalable, as new nodes can quickly join the swarm, contributing additional computational resources and knowledge. This makes it well-suited for applications like medicine and healthcare, where the amount of data or computational requirements may vary between different hospitals.

Swarm learning leverages the concept of distributed learning, where the overall model is built through the collaboration of multiple nodes. Since a SL setting allows many centers to train a model together while keeping training data decentralized, it can protect privacy-sensitive medical data. In the concluding phase of this work, we will get deep into the dynamics of swarm learning behavior by systematically exploring various data configurations between the nodes. This investigation analyzes how swarm intelligence adapts and performs across diverse scenarios.

1.2 Scientific proposal

This work will explore these three critical aspects of artificial intelligence for medicine. The first part of this work will go deep into the field of explainability and reproducibility of SHAP. In the second part, the focus will be on the scarcity and privacy of data in the medical field, introducing the concept of swarm learning.

In Chapter 2, three tools for explainability used for tabular data are introduced: DeepLIFT, LIME, and SHAP. In Chapter 3, the main validation methods used in machine learning to validate models and performances are explained. In Chapter 4, a new method for reproducing consistent and reproducible explanations of models is proposed, implementing SHAP values in validation techniques. In Chapter 5, two works are presented, in which our approach was implemented to obtain representative and reproducible explanations in predicting autism disorder and dementia transition using machine learning. In Chapter 6, the concept of *swarm learning* is introduced, a recent machine learning technique to overcome privacy issues in artificial intelligence for medicine, and its performances in different scenarios are investigated.

Chapter 2

Explainable artificial intelligence

2.1 Explainability

Explainability is the concept that a machine learning model and its output can be explained in a way that "makes sense" to a human being at an acceptable level. It can be formally defined as follows [Chazette et al., 2021]:

> *A system S is explainable with respect to an aspect X of S relative to an addressee A in context C if and only if there is an entity E (the explainer) who, by giving a corpus of information I (the explanation of X), enables A to understand X of S in C.*

The explainability of AI algorithms is becoming more and more critical in many fields of research [Miller, 2019, Cirillo et al., 2020, Arrieta et al., 2020], especially in Medicine. Clinicians must trust algorithms when a prediction occurs and make crucial decisions that can have physical and psychological implications for patients. Therefore, an explainable model should clarify how it arrived at a specific decision and its relevant features. In Medicine, explainability is thus necessary for AI to ensure concordance with medical goals [Adadi and Berrada, 2018].

In our quest for transparent and explainable machine learning models, we will present in the next chapters a new algorithm that leverages representative SHAP values to enhance model explainability. Before delving into this new approach, we will precede the introduction of our new algorithm by exploring and incorporating various other state-of-the-art explainability techniques. Let us introduce the first two explanability methods, DeepLIFT and LIME.

2.2 DeepLIFT

Learning Important Features Through Propagating Activation Differences, or DeepLIFT, is an explainability method that decomposes the output prediction of a neural network on a specific input by backpropagating the contributions of all neurons in the network to every feature of the input [Shrikumar et al., 2017]. This technique works only with Deep learning models, such as artificial neural network (ANN). Indeed, it propagates an essential signal from an output neuron backward through the layers to the input in one pass, like similar backpropagation approaches [Simonyan et al., 2013, Selvaraju et al., 2017]. DeepLIFT explains the difference in output from some reference output in terms of the difference of the input from some reference input [Shrikumar et al., 2017]. This reference input is an input image which is used to explain the input pixel: The reference input represents some default or neutral input chosen according to what is appropriate for the problem. Let us represent some target output neuron of intereset and let $x_1, x_2, \ldots, x_n$ represent some neurons in some intermediate layer. Let y_0 represent the reference activation function y. We define the quantity Δy as the difference from reference: $\Delta y = y - y_0$. DeepLIFT assigns contribution scores $C_{\Delta x_i \Delta y}$ to Δx_i such that it satisfies the first axiom:

- **Axiom 1. Conservation of Total Relevance:** Sum of relevance of all inputs must equal the difference between the score of the input image and the baseline image, at every neuron: Given a reference input vector $\mathbf{x_0}$ with score y_0 and an input vector $\mathbf{x}$ with score y, we define:

$$\Delta \mathbf{x} = \mathbf{x} - \mathbf{x_0} \tag{2.1}$$
$$\Delta y = y - y_0 \tag{2.2}$$

  The definition of the Contribution is given by:

$$\sum_{i=0}^{n} C_{\Delta x_i \Delta y_i} = \Delta y \tag{2.3}$$

$C_{\Delta x_i \Delta y_i}$ can be considered as the amount of difference-from-reference in y that is attributed to the difference-from-reference of x_i. Furthermore, $C_{\Delta x_i \Delta y_i}$ can be non-zero even when $\frac{\partial y}{\partial x_i}$ is zero. This allows DeepLIFT to address a limitation of gradients, which cause noisy attribution maps. For a given input neuron $\mathbf{x}$ and target neuron y, we define the multiplier $m_{\Delta x_i \Delta y}$ following the second axiom:

- **Axiom 2. Back Propagation/Chain Rule:** The contribution per input follows the chain rule like gradients. The definition of contribution per input is given by:

$$m_{\Delta x_i \Delta y_i} = \frac{C_{\Delta x_i \Delta y_i}}{\Delta x_i} \tag{2.4}$$

and following the chain rule we obtain:

$$m_{\Delta x_i \Delta y_i} = \sum_{i=0}^{n} m_{\Delta x \Delta y_i} m_{\Delta y_i \Delta z} \tag{2.5}$$

So the multiplier $m_{\Delta x_i \Delta y_i}$ is the contribution of Δx_i to Δy_i divided by Δx_i.

However, it can be essential to treat positive and negative contributions differently, so let us consider a neuron y with inputs $\mathbf{x}$ such that $y = f(x_1, x_2, ...x_n)$. For every neuron y , we introduce Δy^+ and Δy^- to represent the positive and negative components of Δy, such that:

$$\Delta y = \Delta y^+ + \Delta y^- \tag{2.6}$$

$$C_{\Delta x \Delta y} = C_{\Delta x \Delta y^+} + C_{\Delta x \Delta y^-} \tag{2.7}$$

In the paper, the authors present two different rules for assigning scores:

- **Rescale rule:** this rule applies nonlinear transformations that take a single input such as rectified linear unit (ReLU) [Agarap, 2018] or sigmoid functions, so we set Δy^+ and Δy^- proportional to Δx^+ and Δx^- as follows:

$$\Delta y^+ = \frac{\Delta y}{\Delta x} \Delta x^+ = C_{\Delta x^+ \Delta y^+} \tag{2.8}$$

$$\Delta y^- = \frac{\Delta y}{\Delta x} \Delta x^- = C_{\Delta x^- \Delta y^-} \tag{2.9}$$

Ande we get:

$$m_{\Delta x^+ \Delta y^+} = m_{\Delta x^- \Delta y^-} = m_{\Delta x \Delta y} = \frac{\Delta y}{\Delta x} \tag{2.10}$$

- **RevealCancel rule:** this rule treats the positive and the negative contributions separately. Instead of assuming that Δy^+ and Δy^- are proportional to Δx^+ and Δx^- and that $m_{\Delta x^- \Delta y^-} = m_{\Delta x^+ \Delta y^+} = m_{\Delta x \Delta y}$ as is done for the Rescale rule, we define them as:

$$\Delta y^+ = \frac{1}{2}(f(x^0 + \Delta x^+) - f(x^0)) + \frac{1}{2}(f(x^0 + \Delta x^- + \Delta x^+) - f(x^0 + \Delta x^-)) \tag{2.11}$$

$$\Delta y^- = \frac{1}{2}(f(x^0 + \Delta x^-) - f(x^0)) + \frac{1}{2}(f(x^0 + \Delta x^+ + \Delta x^-) - f(x^0 + \Delta x^+)) \tag{2.12}$$

$$m_{\Delta x^+ \Delta y^+} = \frac{C_{\Delta x^+ \Delta y^+}}{\Delta x^+} = \frac{\Delta y^+}{\Delta x^+} \tag{2.13}$$

$$m_{\Delta x^- \Delta y^-} = \frac{C_{\Delta x^- \Delta y^-}}{\Delta x^-} = \frac{\Delta y^-}{\Delta x^-} \tag{2.14}$$

where x^0 is the reference activation of the input.

DeepLIFT works by comparing the activations of neurons on the actual input to the activations of the neurons on a "reference" or "baseline" input and backpropagating an importance signal ("contribution scores") in such a way that the sum of contributions across all input features will equal the difference of the output activation from its reference value. Using the difference-from-reference allows information to propagate even when the gradient is zero. Essentially, DeepLIFT digs back into the feature selection of the neural network and finds neurons and weights that significantly affect output formation. It gives separate consideration to positive and negative contributions. However, DeepLIFT is model-specific because it is designed specifically for deep neural networks, more specifically Keras and TensorFlow models; it is not compatible with other machine learning models that are not based on neural networks, like extreme gradient boosting (XGBoost) [Chen and Guestrin, 2016a] or support vector machines (SVM) [Cortes and Vapnik, 1995].

2.3 Local Interpretable Model-agnostic Explanations (LIME)

Local Interpretable Model-agnostic Explanations (LIME) is an explainability framework that enables the generation of local explanations of black-box models performing *superpixel occlusion*. LIME was developed by the University of Washington researchers [Ribeiro et al., 2016] to study what happens inside an algorithm by capturing feature interactions. This technique generates the explanations by approximating the model by an interpretable one, such as a linear model, based on the perturbations of the original model. It performs various multi-feature perturbations around a particular prediction and measures the results. The critical aspect behind LIME is that it is much easier to approximate a model by a simple model locally, in the neighborhood of the prediction we want to explain, instead of trying to resemble a model globally.

We define a vector $\mathbf{x} \in \mathbb{R}^{\mathbf{d}}$ to be the original representation of an instance being explained, and we define $\mathbf{x}' \in \{0,1\}^d$ to denote a binary vector for its interpretable representation. For image classification, an interpretable representation may be a binary vector indicating the presence or the absence of a contiguous patch of similar pixels. We define an explanation as a model $g \in G$ where $G \in \{0,1\}^d$ is a class of potentially interpretable models, such as linear models, where g acts over the absence or presence of interpretable components. As not every explanation may be simple enough to be interpretable, let $\Omega(g)$ be a measure of complexity of the explanation $g \in G$. For example for linear models, $\Omega(g)$ may be the number of non-zero weights. Let the model being explained be $f : \mathbb{R}^d \to \mathbb{R}$. We further use $\pi_{\mathbf{x}}$ as a proximity measure between an instance $\mathbf{z}$ to $\mathbf{x}$, so as to define locality around x. Also, let $\mathcal{L}(g, f, \pi_{\mathbf{x}})$ be a measure of how unfaithful g is in approximating f in the locality defined by $\pi_{\mathbf{x}}$. In order to ensure both interpretability and local fidelty, we minimize $\mathcal{L}(g, f, \pi_{\mathbf{x}})$ while having $\Omega(g)$ be low enough to be interpretable by humans. The explanations produced by LIME is obtained by the following equation:

$$\xi(\mathbf{x}) = \operatorname{argmin} \mathcal{L}(g, f, \pi_{\mathbf{x}}) + \Omega(g) \tag{2.15}$$

Presenting the explanation as an optimization problem to find a trade-off between the local fidelity of the explanation and its interpretability it offers no guarantees that the explanations are faithful and stable. Using neighborhood-around explanation instances, it may fall into a curse of dimensionality trap. To overcome this issue, an idea worth pursuing is an axiomatic approach based on the Shapley values [Shapley, 1952]. They

are classic game theory solutions for the distribution of credits to players participating in a cooperative game. Such an approach addresses another limitation of LIME: the ordering of variables impacts the contributions calculated, especially for non-additive models. Shapley-based approaches reduce this issue by averaging the value of a variable's contribution or a large number of possible orderings. Shapley values are a particular example of perturbation-based methods, known as the unique methods that satisfy specific properties, where no hyperparameters are required.

2.4 SHapley Additive exPlanations (SHAP)

Shapley values

Shapley values are a concept taken from the *cooperative game theory* and are used to attribute a player's *contribution* to the result of a game [Shapley, 1952]. Let us consider a cooperative game where a set of players 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 result. Now, let us imagine a machine learning model as a game in which features cooperate to produce a model output and associate each feature with a contribution to 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 to re-train 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 as $f_{S\cup\{i\}}$ a model trained with the i-th feature present and f_S a model trained without the i-th feature. 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 feature i-th 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:

$$\phi_i = \frac{1}{|F|} \sum_{S\subseteq F\setminus\{i\}} \binom{|F|-1}{|S|}^{-1} [f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)] \quad , \tag{2.16}$$

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 $|S|$. Also, we have to divide them by the number of features participating in the model prediction, i.e., 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 [Lundberg and Lee, 2017b]:

$$g(z) = \phi_0 + \sum_{i=1}^{M} \phi_i z_i \quad , \tag{2.17}$$

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. For Lundberg and Lee there are three favorable properties that an explanation model should satisfy [Lundberg and Lee, 2017b]:

> **local accuracy**: the explanation model $g(x')$ matches the model output $f(x)$ when x' is in the neighborhood of x.
>
> **consistency**: if a model changes so that the marginal contribution of a feature value increases or stays the same (regardless of other features), the Shapley value also increases or stays the same.
>
> **missingness**: a missing feature does not contribute to the explanation of the model output.

Shapley values satisfy the first two of these properties, but still cannot handle missing values. In order to satisfy also the missingness property, Lundberg and Lee proposed the SHAP values, explained in the next lines.

SHAP values

SHAP values are the classical Shapley values of a *conditional expectation function* of the original model f [Lundberg and Lee, 2017b]. 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 [Lundberg and Lee, 2017b]. For this reason, SHAP values provide the 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 in 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 feature-less model, i.e., the average of the samples' outcomes $E[f(z)]$.

The SHAP method is preferable to other explanation methods because it satisfies the properties of local accuracy, consistency, and missingness [Lundberg and Lee, 2017b]. A comparison between SHAP and other XAI methods (e.g., LIME) has already been explored in the literature (see, e.g., [Lombardi et al., 2021, Antwarg et al., 2021]). In particular, Lombardi et al. [Lombardi et al., 2021] showed that SHAP values can provide more reliable explanations, i.e., less influenced by small variations of the training set. However, since SHAP depends on inherently stochastic model predictions, the explanations might exhibit variability in different training iterations. Moreover, SHAP framework has been proposed for hold-out strategies [Batunacun et al., 2021, Bi et al., 2020b, Kim and Kim, 2022, Chen et al., 2019, Rodríguez-Pérez and Bajorath, 2020], where SHAP values are computed only when a final model is trained. For this reason, extending its application across diverse validation scenarios is essential. Some authors adopted the SHAP method with CV strategies [Parsa et al., 2020, Bi et al., 2020a, Feng et al., 2021, Deb and Smith, 2021, Wang et al., 2021, El-Sappagh et al., 2021], but SHAP was used only after the CV procedure on often unclear portions of the dataset. To overcome this issue, in Chapter 4, we propose a method that enables us to get consistent explanations of trained machine learning models in a repeated nested cross-validation procedure. By incorporating SHAP into different validation strategies and conducting multiple repetitions, we aim to ensure that the explanations generated remain consistent and robust across various training instances of the same model. This approach allows for a thorough examination of the algorithm's generalizability across diverse datasets. However, it's important to note that our algorithm is versatile and adaptable to various validation methods beyond the presented framework. Recognizing the significance of employing appropriate validation techniques, the next chapter will comprehensively introduce and discuss the main validation methods suitable for our algorithm.

Chapter 3

Validation Methods

This Chapter introduces the concept of *validation* in machine learning. Model validation is a core component of developing machine learning or artificial intelligence that assesses the ability of an ML or statistical model to produce predictions with enough accuracy and performance. It is an essential step in developing any ML or AI system, as it helps ensure that the model performs as intended and can handle unseen data. With proper model validation, the confidence in its ability to generalize well on unseen data can be high. Furthermore, validation helps determine the best model, parameters, and accuracy metric for the given problem.

Model validation also allows for comparing different models, allowing us to choose the best one for the task. Furthermore, it helps determine the model's accuracy when presented with new data.

3.1 Holdout Validation

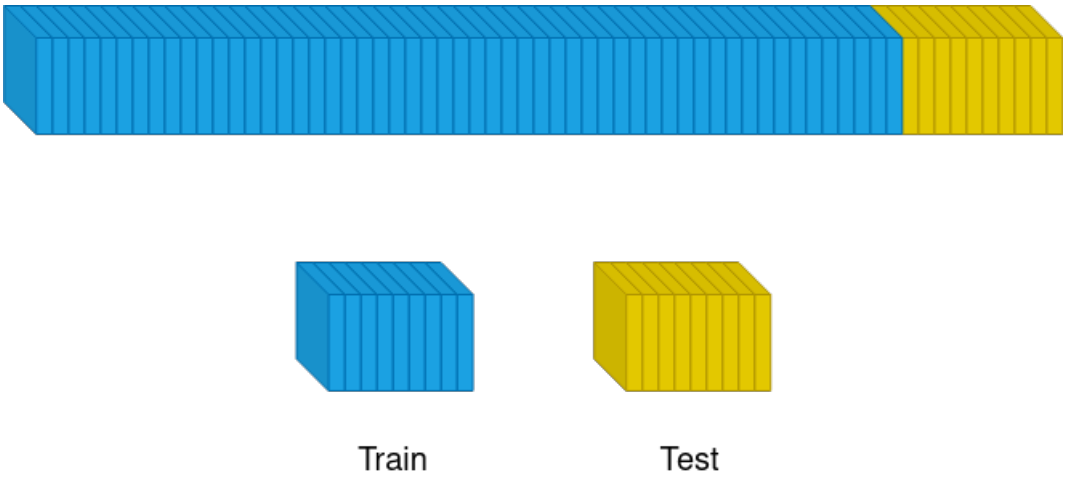


Figure 3.1: Example of hold-out method. The original dataset is divided in training set and test set. The training set is used to train the model and the test set is used to assess the model performance with unseen data.

Holdout validation is the most common approach for evaluating ML models. In this approach, the available data are partitioned into *training*, *validation*, and *test* sets. The proportion of the available data used for each set depends on the number of available data points, the data variability, and the characteristics of the used model. Generally, the proportion of the validation set assigned to the training data must be significant when working with small datasets. Typically, 70% of the available data are used for training, 15% for validation, and the remaining 15% for testing the model, although the percentage allocations can vary.

Due to the intrinsic size of validation sets and better representation of the data, the ratio between validation and training sets can be smaller as datasets grow. The training and validation sets are used to build the models. Model parameters are learned from the training set. Model hyperparameters are determined using the validation data during a process called *hyperparameter tuning*. Models are trained, tuned and then evaluated on the test set to estimate their generalization error (i.e. the error of the resulting model when applied to unseen data). Training and tuning the models should not use the test data; otherwise, the generalization error estimate would be overoptimistic and inaccurate. Because of its computational efficiency, the holdout validation strategy is frequently employed for training

deep learning models with big datasets. However, this method is often criticized, for not utilizing the entire dataset for small datasets. It is possible that a tiny test set will not yield a trustworthy estimate of model performance, and that the performance measures depend on the selection of the test. It is frequently impossible to choose a test set big enough to be reflective of the underlying data for small datasets. Furthermore, fewer samples are available to train the model when a bigger test set is employed, which has an adverse effect on the final model's performance.

Also, when fine-tuning a model using this approach, the resulting model may be sensitive to the choice of the validation set, resulting in models with low ability to generalize.

3.2 Cross-Validation

Cross-Validation (CV) is a resampling approach used to evaluate ML models. This method provides an unbiased estimation of model performance [Hastie, 2013]. This approach tends to give a more accurate estimate of generalization error when dealing with small datasets than hold-out validations. In the following section, we introduce different types of cross-validation.

K-fold cross-validation

In k-fold cross-validation, samples are divided into k non-overlapping splits (or *folds*) (Figure 3.2). Next, the model is trained k times, and each time, the model is trained on the $k-1$ folds and validated on the remaining k_{th} fold. This way, each of the different k folds is selected only once to be the validation set. Then, the average performance obtained by the k various trainings of the model is used as the estimate of the validation error. This technique requires much more computational time than hold-out strategies because it includes k different trainings of the model. However, the variance of the performance measure is reduced, and the resulting estimate is more consistent than one single hold-out validation method.

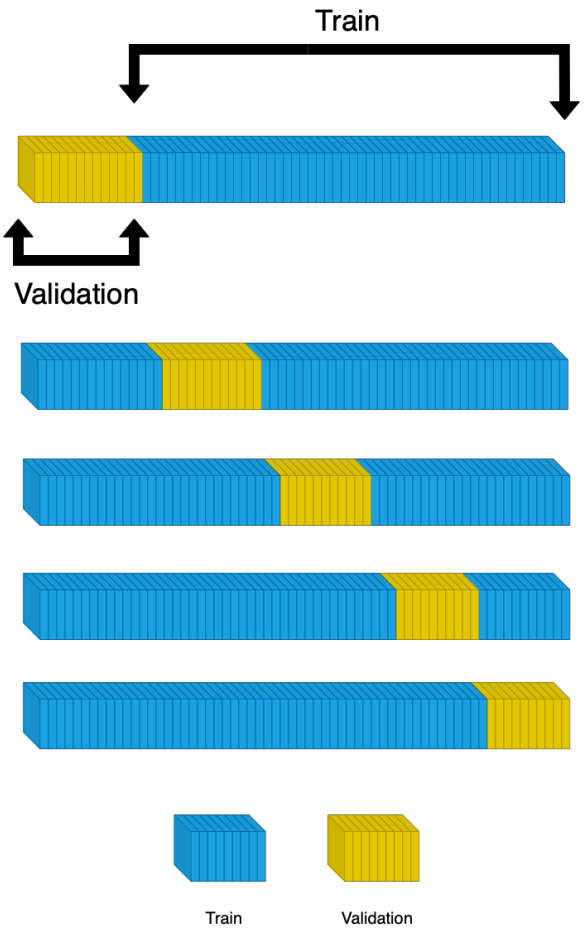


Figure 3.2: k-fold Cross validation with $k = 5$.

The choice of the number of folds k depends on the computational time and the number of samples of the whole dataset. However, 10-fold and 5-fold CV is the most widely used for evaluating ML models [Hastie, 2013]. When dealing with unbalanced datasets, namely datasets where one class of samples is much more frequent than the others, k-fold CV may lead to unstable performance measures. In the *stratified* k-fold CV, each k folds are sampled so that the distribution of the classes in each fold is almost the same in the whole dataset.

Nested cross-validation

Most ML models use several hyperparameters. When developing ML solutions, it is standard practice to tune these hyperparameters. Conducting

experiments to find the hyperparameter values that yield the best results is standard practice. When using multiple models to find the best hyperparameter values for a CV, which divides data into training and validation sets, the validation error that results is frequently over-optimistic when used to estimate generalization error. Consequently, a test set should be kept private and out of the model training and hyperparameter tweaking processes. A trustworthy estimate of generalization error can be obtained from the model's performance on this test set. For small datasets, choosing a single subset of data as the test set yields estimates for generalization errors that have high variance and are sensitive to the composition of the test set. Nested cross-validation (nCV) is used to address this challenge. Indeed, nCV helps examine the unbiased generalization performance of the trained models and simultaneously performs hyperparameters optimization [Mueller and Guido, 2017]. NCV consists of an outer cross-validation loop and an inner cross-validation loop. The outer loop uses different train, validation, and test splits (Figure 3.3). The inner loop takes a train and validation set chosen by the outer loop. The model with different hyperparameters is trained using the training set, and the best hyperparameters are determined based on the performance of the trained models on the validation set. In the outer loop, generalization error is estimated by averaging test error over the test sets in the outer loop.

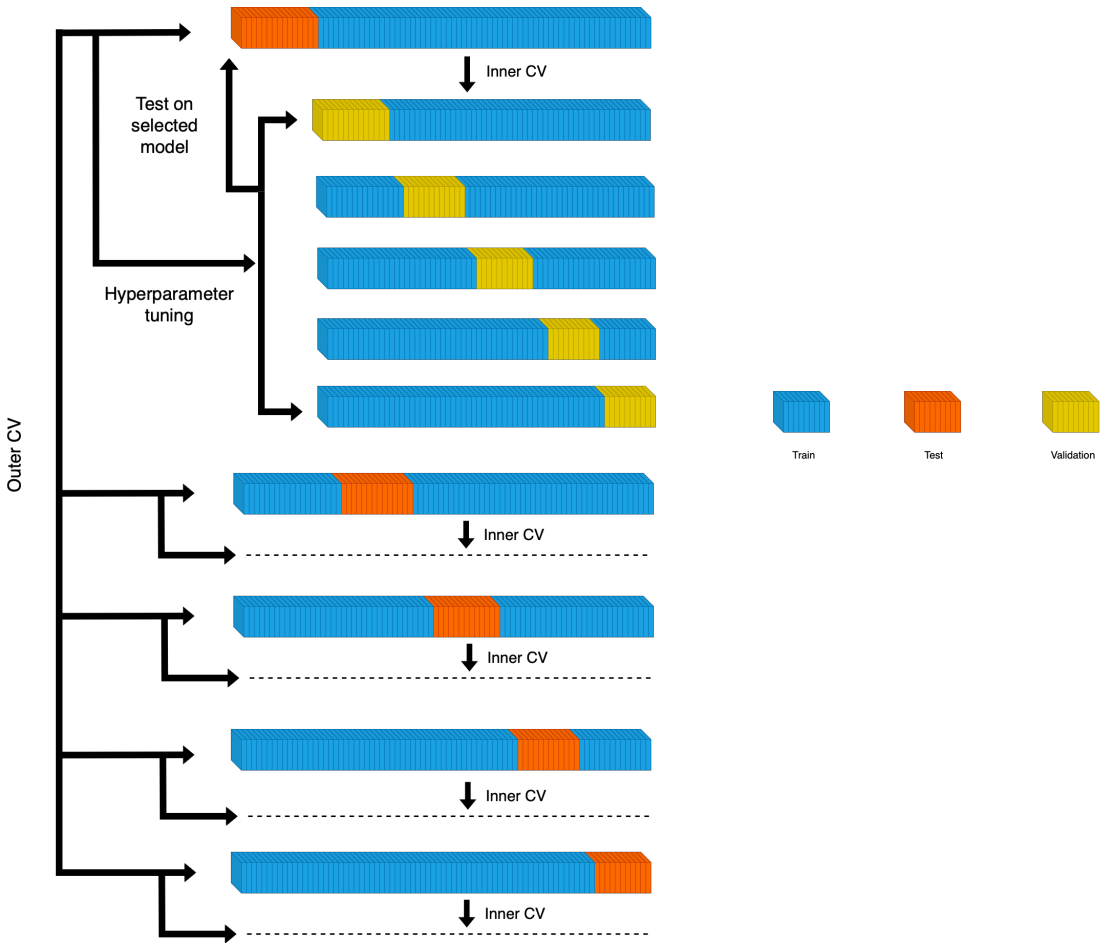


Figure 3.3: Nested Cross validation with 5 folds for the outer loop and 5 folds for the inner loop.

3.3 Boostrap Resampling

The bootstrap method is a resampling technique used to estimate statistics on a population by sampling a dataset with replacement [Dixon, 2006].

It can estimate summary statistics, such as the mean or standard deviation.

A desirable property of the results from estimating ML model performance is that the estimated value can be presented with confidence intervals, a feature not available with other methods such as cross-validation. This method draws a sample of size n from the whole dataset D. Then, the sampling distribution is created by resampling observations with replacement from D m times, with each resampled set having n observations.

Therefore, by resampling the D dataset m times, it would be as if m samples were drawn from the original population, and the estimates derived would be representative of the theoretical distribution under the traditional approach (Figure 3.4).

Increasing the resamples m will not improve the data's information. The amount of information within the set depends on the sample size n, which will remain constant throughout each resample.

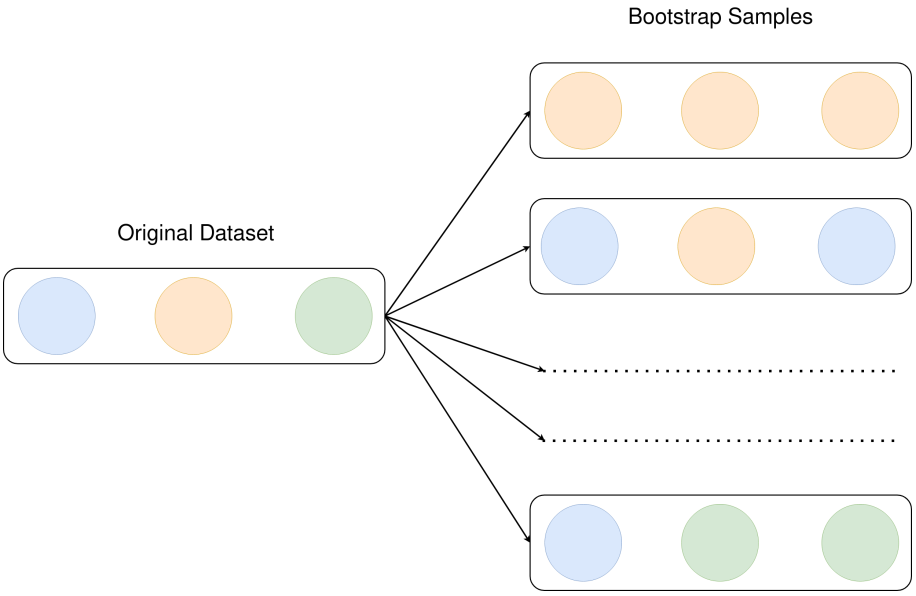


Figure 3.4: Bootstrap resampling.

3.4 Reproducibility and Explainability

Reproducibility is a key concept not only in artificial intelligence research but also in science in general: a scientific experiment must replicate the same results given the same initial conditions every time it is performed. In the same way, ML models need to provide consistent and robust results when performed multiple times. In the Machines Learning field, Reproducibility can be defined as [Gundersen and Kjensmo, 2018]:

> *Reproducibility in empirical AI research is the ability of an independent research team to produce the same results using the same AI method based on the documentation made by the original research team.*

This concept is particularly valid in artificial intelligence for medicine and related fields, where clinicians and patients need consistent and trustworthy AI suggestions for diagnosis and treatments. The problem of reproducibility in science has been debated for the past few decades, especially in Medicine and Healthcare [Beam et al., 2020, Rajpurkar et al., , Haibe-Kains et al., , Stower, 2020, Walsh et al., 2021].

The main problem of this concept is the fact that one of the main characteristics of many machine learning strategies is their inherent *stochastic* nature [Sabuncu, 2020], which leads the performance not to be precisely reproducible. For example, even the choice of random seeds in many machine learning models (whenever they are present) could lead to the high variability of the performance between two different training procedures of the same model [Amir et al., 2021]. 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., [Li et al., 2020, Kim, 2009, Burman, 1989, Vanwinckelen and Blockeel, 2012]). Furthermore, due to the frequent scarcity of data in medical research [Lee Choong Ho, 2017], one approach is the repetition of a nested cross-validation loop [Mueller and Guido, 2017]. This method is especially effective when the data is relatively low because it allows for training and testing a model, often using broad, non-overlapping folds of the dataset. This approach allows the possibility of testing the model on all dataset subjects and obtaining an average performance. These fundamental concepts also have to be transferred to Explainability. A given explanation of a model has to be consistent and the same explanation for every repetition of a machine learning experiment. It is imperative that the explanations provided by a model remain

consistent and reliable across multiple repetitions of an experiment. Furthermore, in scenarios of limited data availability, particularly common in medical research, the repetition of a nested cross-validation loop emerges as a valuable approach. This method, discussed in the next chapter, proves effective in training and testing models across diverse, non-overlapping folds of the dataset, contributing to a more robust and reliable evaluation of model performance. In the upcoming discussion, we will delve into the application of SHAP values within a repeated nested validation setting, shedding light on their role in enhancing both performance and explainability.

Chapter 4

Representative SHAP values

The study reported in this chapter refers to the published journal paper entitled *"Explanations of Machine Learning Models in Repeated Nested Cross-Validation: An Application in Age Prediction Using Brain Complexity Features"*, Scheda R., Diciotti S.

In the previous chapters, the critical role of reproducibility and explainability discussed. It has been shown how these concepts are integral components of a robust machine learning framework, emphasizing the crucial requirement that explainability must possess the attribute of being reproducible. In this work, it's proposed 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. This method was applied to predict individual age using brain complexity features extracted from Magnetic Resonance Imaging (MRI) scans of 159 healthy subjects. In conclusion, this proposed method allows a rigorous assessment of the SHAP explanations of a trained model in a repeated nested cross-validation setting.

4.1 Introduction

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 [Shapley, 1952]. As mentioned in Chapter 2, despite other explainability methods, SHAP is a preferable explainability technique because it satisfies the properties of *local accuracy*, *missingness*, and *consistency*. SHAP method is also preferable because it allows us to get global explanations but also sample-level explanations of a machine learning model [Lombardi et al., 2021].

Nevertheless, 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, SHAP framework has been proposed for hold-out strategies [Batunacun et al., 2021, Bi et al., 2020b, Kim and Kim, 2022, Chen et al., 2019, Rodríguez-Pérez and Bajorath, 2020], where SHAP values are computed only when a final model is trained.

To make explanations consistent and reproducible, it becomes necessary to integrate SHAP values computation within different validation techniques such as repeated CV and repeated bootstrap resampling. By embedding SHAP values computations within these validation frameworks, we ensure a better understanding of feature contributions that are sensitive to the model's predictive capabilities and resilient to potential biases or variations in the training data. For these reasons, in this study, we extended the use of SHAP values for a repeated nCV setting by estimating *representative* SHAP values, obtained by averaging patient-level SHAP values of each feature across different folds and repetition of the nested CV. Moreover, we consider it essential to evaluate the SHAP values separately for the training and test set of the different cross-validation rounds to evaluate the generalization abilities of the SHAP explanations of a trained model. To test our method, we applied representative SHAP values computed in nCV to two regression and one classification task in predicting individual age using brain complexity features from two public, and international neuroimaging datasets of in-vivo MRI scans for a total of 159 healthy subjects (age range 6–85 years).

4.1.1 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 4.1).

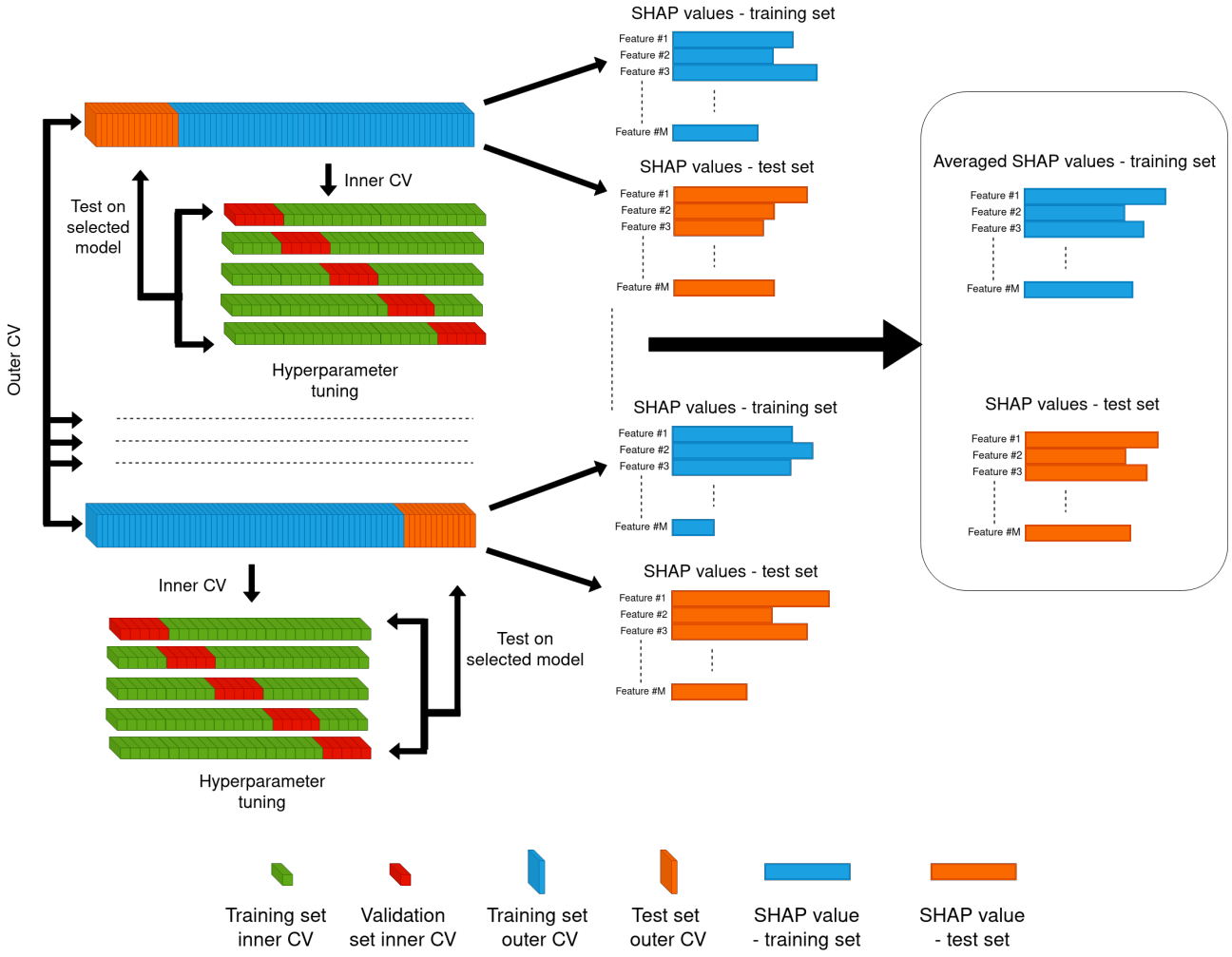


Figure 4.1: Schematic representation of the computation of representative SHAP values. SHAP values of training and test folds are computed separately for each round of the outer CV. Then, SHAP values are averaged over the training folds. This procedure is repeated K times, and SHAP values for the training and test sets are averaged over the R repetitions. (Image adapted from [Yagis et al., 2021].)

This strategy involves nesting two K-fold CV loops where the inner loop is used to optimize, e.g., model hyperparameters. The outer loop gives an unbiased estimate of the performance of the best model [Mueller

and Guido, 2017]. 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, i.e., $K-1$ 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* [Lundberg, 2018c]. 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 the $K-1$ folds of the outer CV as follows:

$$(\phi_{train})_n^{ir} = \frac{1}{K-1} \sum_{k=1}^{K-1} \left(\phi_{nk}^{ir} \right) \quad . \tag{4.1}$$

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

$$(\phi_{test})_n^{ir} = \phi_{nk^*}^{ir} \quad . \tag{4.2}$$

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

$$(\bar{\phi}_{train})_n^{i} = \frac{1}{R} \sum_{r=1}^{R} (\phi_{train})_n^{ir} \quad , \tag{4.3}$$

$$(\bar{\phi}_{test})_n^{i} = \frac{1}{R} \sum_{r=1}^{R} (\phi_{test})_n^{ir} \quad . \tag{4.4}$$

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

$X \leftarrow Dataset$ # *Data table with N samples and M features (N rows × M columns)*
$y \leftarrow Target$
$model \leftarrow Classifier$
$Explainer \leftarrow SHAP.Explainer()$ # *Shap function which computes SHAP values*
$train\_folds\_shap\_values \leftarrow 0$ # *Initialized matrix (N rows × M columns)*
$test\_folds\_shap\_values \leftarrow 0$ # *Initialized matrix (N rows × M columns)*
for r `in` $1, \ldots, R$ **do**
  $fold\_splits \leftarrow split(K)$
  $innerCV(fold\_splits, model)$
  $outerCV(innerCV, K).fit(X, y)$
  for k `in` $fold\_splits$ **do**
    $X\_train, y\_train \leftarrow fold\_splits.train(k)$
    $X\_test, y\_test \leftarrow fold\_splits.test(k)$
    $best\_k\_model \leftarrow outer\_CV[k].best\_model$
    $best\_k\_model.fit(X\_train, y\_train)$
    $train\_shap\_values \leftarrow Explainer(best\_k\_model(X\_train))$
    $test\_shap\_values \leftarrow Explainer(best\_k\_model(X\_test))$
    $train\_folds\_shap\_values \leftarrow train\_folds\_shap\_values + \frac{train\_shap\_values}{K-1}$
    $test\_folds\_shap\_values \leftarrow test\_folds\_shap\_values + test\_shap\_values$
  end for
end for
$average\_train\_folds\_shap\_values = train\_folds\_shap\_values / R$ # $\bar{\phi}_{train}$
$average\_test\_folds\_shap\_values = test\_folds\_shap\_values / R$ # $\bar{\phi}_{test}$

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 [Franke and Gaser, 2019]. In this context, biomarkers quantifying brain complexity, including the local gyrification index (lGI) and the fractal dimension (FD) of the cerebral gray and white matter, have been proved to have predictive capabilities in age prediction [Franke and Gaser, 2019, Marzi et al., 2020, Madan and Kensinger, 2016]. 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 [Marzi et al., 2020].

To prove the utility of computing the SHAP values in repeated nCV, in this study, we considered two regression and one classification task for age prediction using features of brain complexity extracted from MRI data [Marzi et al., 2020]. In particular, we used the high-resolution public and international T_1-weigthed datasets of healthy children and adolescents [Nathan Kline Institute (NKI)—Rockland Sample Pediatric Multimodal Imaging Test–Retest Sample—NKI2 dataset [Nooner et al., 2012]] and adults [International Consortium for Brain Mapping (ICBM) dataset

[Kötter et al., 2001]]. 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 ± 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 having age $\leq$ 30 years (25 subjects – 9 males and 16 females, 22.6 ± 3.3 years) and the elder group of subjects having age $\geq$ 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 [Marzi et al., 2020]. 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/`) [Fischl, 2012], a dedicated brain segmentation software [Rosas et al., 2002, Han et al., 2006, Lee et al., 2006, Kang et al., 2012, Keller et al., 2013, King, 2014]. 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 [Fischl et al., 2001] 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 lGI was computed following a surface-based approach [Schaer et al., 2008]. 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 lGI 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 lGI within the entire cortex to obtain a gyrification index (GI) representative of the cortical complexity of each subject. Moreover, we recorded the following FreeSurfer's 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. They include

the use of (1) a priori selection of the interval equal to [4 mm–256 mm] (inspired by Kiselev et al. [Kiselev et al., 2003]) ($FD_{A\ priori\ \#1}$) (2) a priori selection equal to 5–40% of the smallest Euclidean dimension of the cerebral cortex [Goñi et al., 2013] (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 [Marzi et al., 2018, Pantoni et al., 2019a] ($FD_{Auto\ Marzi\ et\ al.\ 2018}$), (4) an improved automated selection of the interval of spatial scales, based on the search of the interval of spatial scales which presents the highest rounded R^2_{adj} coefficient and, in case of equal rounded R^2_{adj} coefficient, preferring the widest interval in the log–log plot ($FD_{Auto\ fractalbrain}$) [Marzi et al., 2020, Marzi et al., 2021a, Pani et al., 2022b].

We predicted individual age using an 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 [Chen and Guestrin, 2016b]. 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 $colsample\_bytree \in \{0.25, 0.5, 0.75, 1\}$, the maximum depth of a tree $max\_depth \in \{2, 3, 4\}$, the minimum number of instances needed to be in each node $min\_child\_weight \in \{2, 3, 5\}$, the number of decision trees $n\_estimators \in \{5, 10, 20, 100\}$ and the ratio of training data randomly sampled prior to growing trees $subsample \in \{0.1, 0.2, 0.4\}$. Moreover, since SHAP has an optimized implementation for tree-based models (called *TreeExplainer*), using XGBoost, we can compute SHAP values in polynomial time compared to model agnostic explainers for this class of models [Lundberg et al., 2020]. We adopted a repeated (100 times) nCV strategy, and we chose a 5-fold CV in both the inner and outer loops because it offers a favorable bias-variance trade-off [Hastie, 2013].

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.

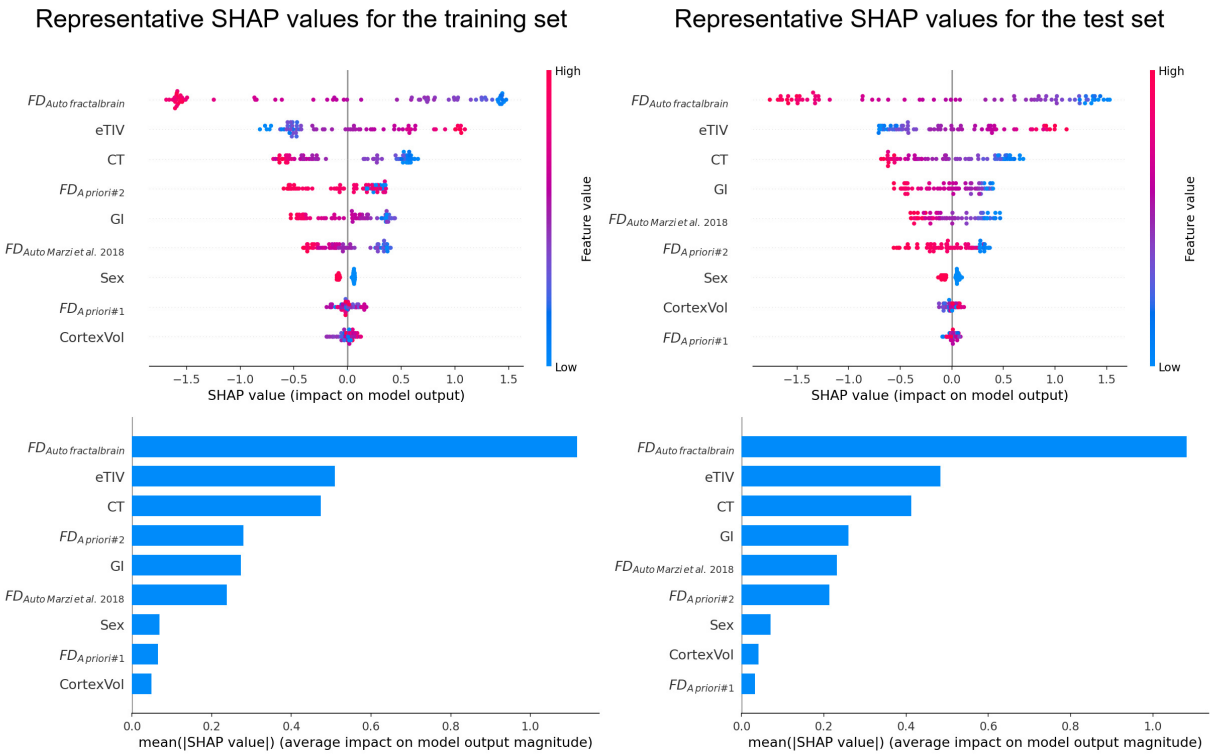


Figure 4.2: Results for the NKI2 regression task in a 5-fold nCV over 100 repetitions. Top row: beeswarm summary plots of representative SHAP values for the training (on the left) and test sets (on the right). The given SHAP explanation is represented by a single dot on each feature row for each sample (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).

4.2 Results

For the two regression tasks, we obtained an MAE of 1.61 $\pm$ 0.14 years (mean $\pm$ standard deviation) in the NKI2 dataset and 12.13 $\pm$ 0.86 years in the ICBM dataset. For the classification task, we obtained a ROC AUC value of 0.881 $\pm$ 0.068 and balanced accuracy of 0.77 $\pm$ 0.06 (Figure 4.5). 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.

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 4.2 and 4.3 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 [Lundberg, 2018b]. 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 [Lundberg, 2018a].

The same plots for the classification task using the ICBM dataset are shown in Figure 4.4.

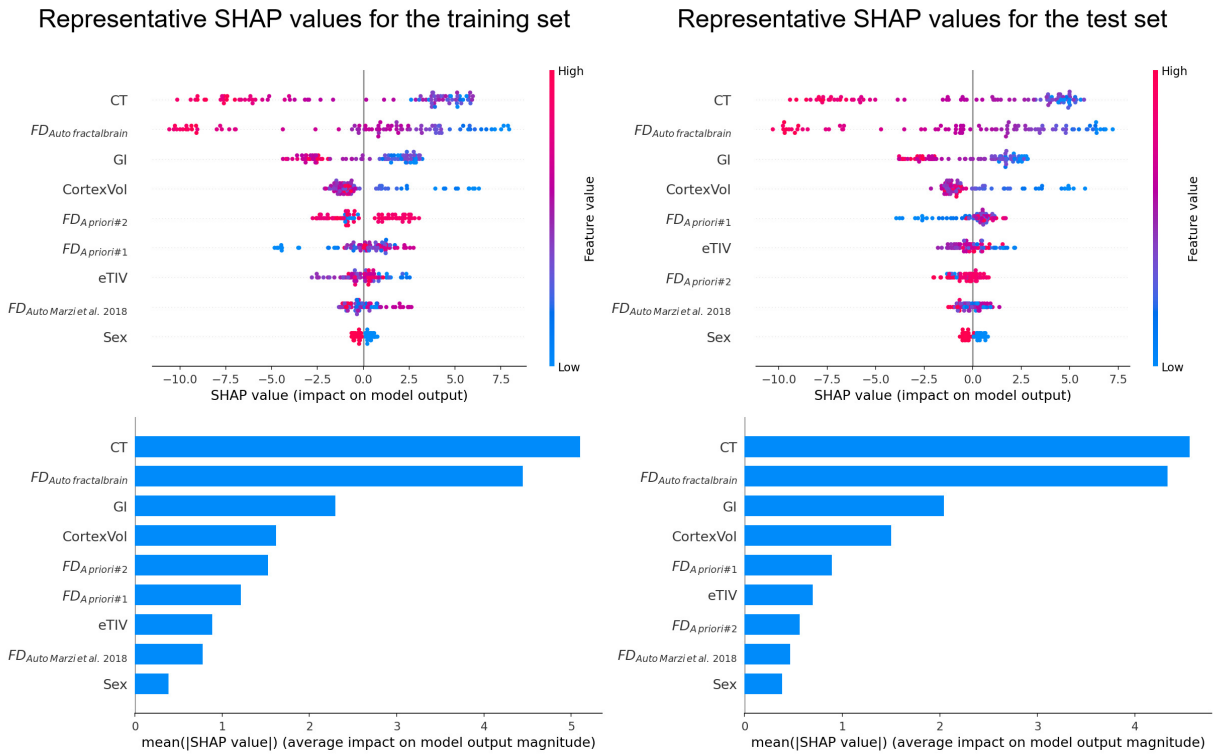


Figure 4.3: Results for the ICBM regression task in a 5-fold nCV over 100 repetitions. Top row: beeswarm summary plots of representative SHAP values for the training (on the left) and test sets (on the right). The given SHAP explanation is represented by a single dot on each feature row for each sample (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).

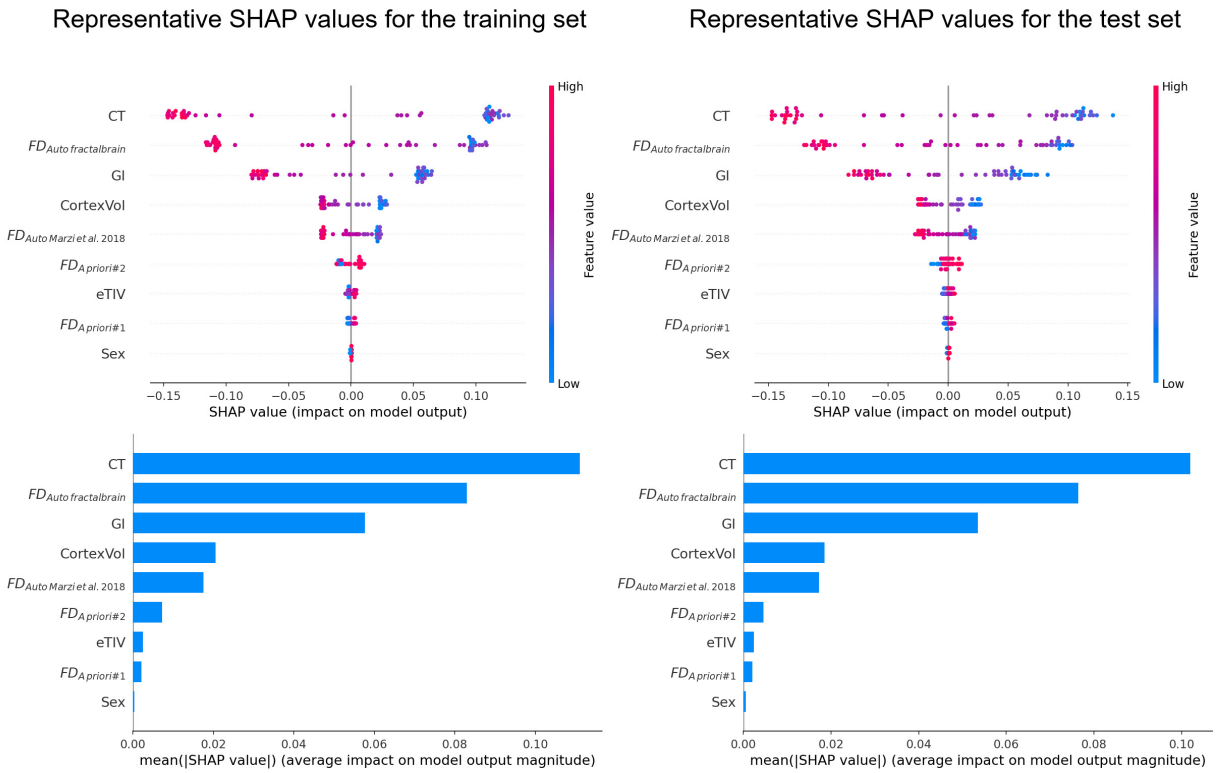


Figure 4.4: Results for the ICBM classification task in a 5-fold nCV over 100 repetitions. Top row: beeswarm summary plots of representative SHAP values for the training (on the left) and test sets (on the right). The given SHAP explanation is represented by a single dot on each feature row for each 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).

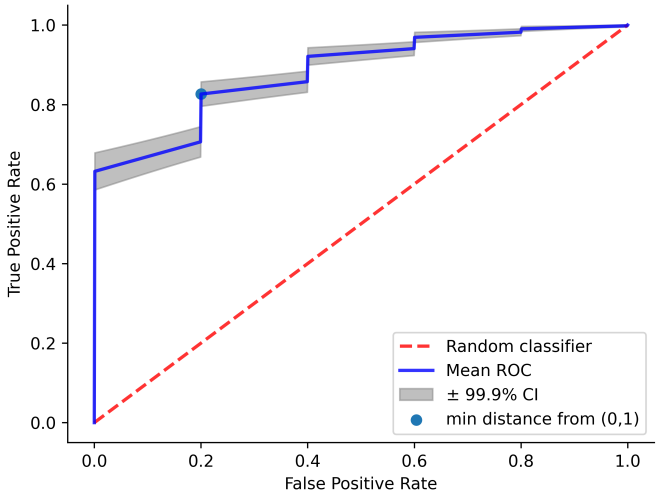


Figure 4.5: Average (and ± 99.9% confidence interval (CI)) ROC curve of the model trained in the classification task in a 5-fold nCV over 100 repetitions using the ICBM dataset. The point in the ROC curve with the minimum distance from the ideal classifier (at coordinates (0,1)) is represented in blue (at coordinates (0.200, 0.826)). The ROC curve of a random classifier is overlayed in red as a reference.

4.3 Discussion

In this study, we proposed a method to compute representative SHAP values in a repeated nested CV procedure. As for the standard performance of machine learning models, in which average performance metrics are usually given, also representative explainable values acting as a final assessment of the behavior of the entire model, are essential. Whereas 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 [Shapley, 1952], 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 4.6 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.

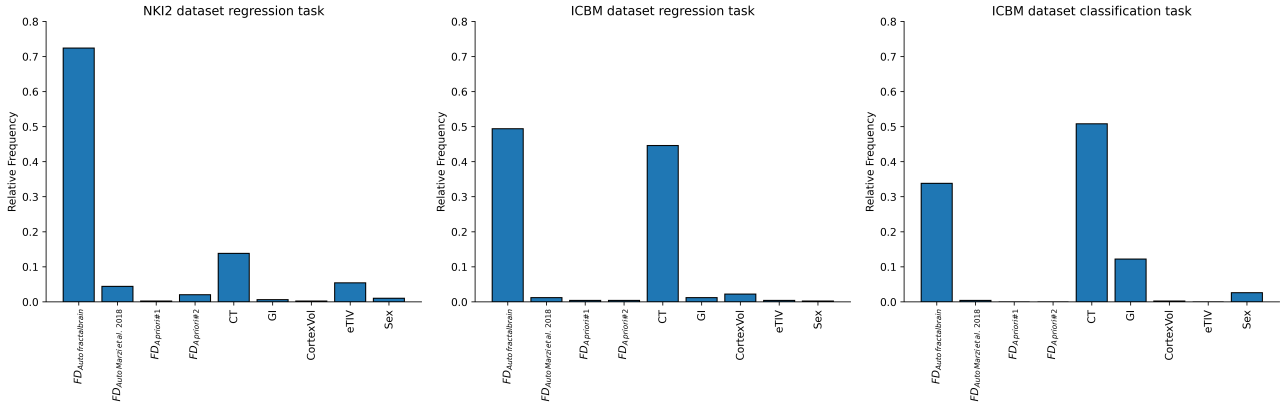


Figure 4.6: Relative frequency of neuroimaging features. For each regression/classification task, the frequency with which each feature was identified as the most important (highest impact as measured by the absolute value of the SHAP value over all the given samples across all outer CV test folds in 100 repetitions is shown.

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, e.g., the summary bar plot representing global feature importance for the regression task using the NKI dataset in Figure 4.7.

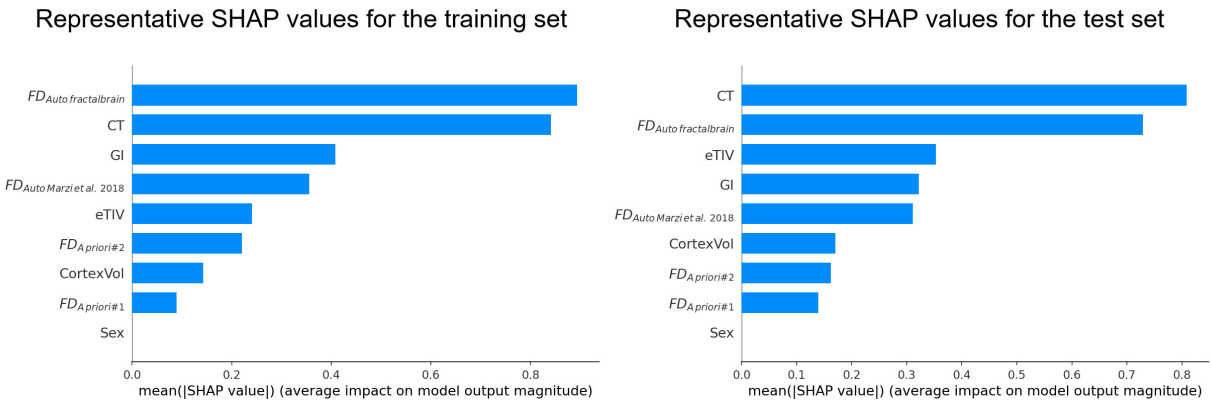


Figure 4.7: Summary bar plot representing global feature importance for a single iteration of the regression task using the NKI dataset, separately, for the training (on the left) and test sets (on the right). The order of the most impactful features in one iteration differs between training and test sets.

Previously, SHAP values were estimated for the test set of a single nCV repetition, thus generating potentially unstable explanations [Beebe-Wang et al., 2021, El-Sappagh et al., 2021, Siciarz et al., 2021]. Blüthgen et al. proposed 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 sign of the SHAP values which inform the positive/negative association with each feature [Blüthgen et al., 2021]. In two recent works [Lombardi et al., 2021, Lombardi et al., 2022], the average of SHAP values of samples in test sets among 100 repetitions of an ML model has been applied. In [Lombardi et al., 2021], 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 under-sampled 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 [Lombardi et al., 2022], the authors adopted a leave-one-subject-out CV strategy using MRI data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, i.e., they split 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 under-sampled 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 multiple times the same model on different portions of training datasets [Lombardi et al., 2021], whereas, in the other study, they added hyperparameter tuning within the repetitions to select the best model given by the different subsets [Lombardi et al., 2022].

Our method allows to estimate representative SHAP values, separately for the training and test sets, in a repeated nCV setting following a well-documented algorithm 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 [Lombardi et al., 2021, Lombardi et al., 2022] for two main reasons: i) we repeat$R$ times the nested CV rather than generating repeated under-sampled training set in a single nested CV, ii) our proposed method allows 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 [Marzi et al., 2020]. We previously showed a monotonic decrease in structural complexity (in terms of $FD_{Auto\,fractalbrain}$) of the cerebral cortex with age during almost all the lifespan [Marzi et al., 2020] and, more recently, that the 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 [Pani et al., 2022b]. 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 4.2, 4.3 and 4.4). In other words, we confirmed that a lower brain complexity and cortical thinning are valuable predictors of older subjects both in a young and adult cohort. Moreover, our results support that our latest development of

the fractal dimension ($FD_{Auto\ fractalbrain}$) [Marzi et al., 2020] is more predictive of individual age than 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, especially fascinating when using ultrahigh-field MRI [Marzi et al., 2021a]. 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 4.3 and 4.4). This result confirms the ability of the fractal dimension of the cerebral cortex, besides that of cortical thickness and gyrification, in characterizing brain maturation and aging, as previously observed for neurodegeneration [King et al., 2010]. 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 4.3), this was not the case for the NKI2 regression task (see Figure 4.2) and ICBM classification task (see Figure 4.4), 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 the tree-based models as the 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, e.g., functional connectivity features extracted by functional MRI [Monti et al., 2020] and electroencephalography (EEG) data [Al Zoubi et al., 2018] that could also potentially use recent deep learning progresses [Zhang et al., 2021, Zhao et al., 2022].

4.4 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, test, and explainability of AI models designed in machine learning pipelines independently of the validation strategy.

Chapter 5

Representative SHAP values in artificial intelligence for neuroscience

In the previous chapter, it was presented a new algorithm designed to compute SHAP values within the context of repeated nested cross-validation. The significance of this algorithm lies in its capacity to enhance the robustness of interpretability of machine learning models in different validation settings. Based on this, this chapter showcases the application of our algorithm in two distinct papers focused on diagnostics in neuroscience. The first paper explores the utilization of our algorithm for the diagnosis of autism spectrum disorder, leveraging tabular data. The second paper concerns the prediction of dementia transition diagnosis, employing tabular data extracted from MRI images. These contributions underscore the algorithm's versatility and practical implications in addressing critical challenges within healthcare, improving understanding, and early detection of neurodevelopmental and neurodegenerative disorders.

5.1 Explainability in autism spectrum disorder diagnosis

The study reported in this section refers to the published journal paper entitled *"Early prediction of Autism Spectrum Disorders through interaction analysis in home videos and explainable artificial intelligence"*, Paolucci, C; Giorgini, F; Scheda, R; Alessi, FV; Diciotti, S.

This work proposes an AI pre-screening tool for early Autism Spectrum Disorders (ASD) diagnosis with the aim of creating an easily administrable tool for observers useful to identify potentially alarming signs in pre-verbal interactions. These features are evaluated using an explainable artificial intelligence algorithm using representative SHAP values to assess which of the proposed new interaction features was more effective in classifying individuals with ASD vs. healthy subjects. We used a rating scale with three core sections: sensorimotor, behavioral, and emotional, each further divided into four sub-features. By seeing home videos of children doing everyday activities, two experienced observers rated each sub-feature from 1, corresponding to typical interactions, to 8 corresponding to extremely atypical interactions. Then, a machine learning model based on XGBoost was developed to identify ASD children. The classification obtained had an area under the receiver operating curve of 0.938 and 0.914 for the two observers, respectively. Representative SHAP values demonstrated the significance of early detection of body-related sensorimotor features.

5.1.1 Introduction

The prevalence of ASD in Europe is 12.2 per 1000 children [Salari et al., 2022]. Early diagnosis of ASD has proven to be crucial in achieving effective treatment [Gabbay-Dizdar et al., 2022], thereby improving the lives of ASD infants and their families [Elder et al., 2017, Franz and Dawson, 2019, Rotholz et al., 2017, van 't Hof et al., 2021, Volkmar, 2014]. This work seeks to develop easily understandable and administrable tools to identify potentially alarming behaviors, which is usually considered a precious achievement by scholars concerned by very early diagnosis [Daniels and Mandell, 2014]. Indeed, evident signs of impairments and atypicality that can lead to ASD can be seen by looking at embodied and prelinguistic interactions between infants and caregivers when the toddler is between 9 and 18 months old [Gallagher and Hutto, 2008, Paolucci, 2020, Trevarthen

and Hubley, 1978]. As several retrospective studies show [Alonim et al., 2021], early symptoms and behaviors related to ASD can be seen long before the infant enters the linguistic phase. It is difficult to provide a diagnosis before the age of two years and a half, which usually follows the observation of a linguistic skills development delay in children [van 't Hof et al., 2021]. Critical issues can also be found within screening tests such as the Autism Diagnostic Observation Schedule (ADOS) [Luyster et al., 2009]. These tests are typically used to screen the general community or a population that is already at risk, identify cases of ASD, and separate them from other conditions that have similar symptoms. This test consists of a series of highly structured activities, during which examiners assess the presence of specific behaviors that are natural to a neurotypical subject and usually lacking or deficient in ASD subjects. Hence, ADOS mostly succeeds for two reasons: i) the observer is a highly competent subject, usually a neuropsychiatrist; ii) ADOS is a highly grammaticalized test carried out in a controlled laboratory situation. However, observers like caregivers, teachers, or parents have difficulty identifying warning signs without any expertise, even if they are the ones who spend most of the time with later ASD-diagnosed infants. Furthermore, the dynamics of daily activities shared between caregivers and infants are not as structured and schematic as those that makeup ADOS tests, with the consequence that even if caregivers had this kind of expertise, it would be difficult for them to apply it in daily, unplanned interactions. This implies that non-expert observers risk not identifying potential warning signs from the earliest months of life. This work tries to overcome this issue, developing easily understandable and administrable tools for teachers, parents or caregivers, the non-competent observers who spend most of the time with future ASD children. To test the validity of this methodology, we used representative SHAP values to evaluate which of the newly proposed interaction features were more effective in classifying individuals with ASD compared to controls.

5.1.2 Materials and methods

Feature extraction

We tried to analytically distinguish the main areas involved in a caregiver-infant interaction through a simplified system in which the observer only needs to look at the attunement between the child and the caregiver during their interaction. Thus, our idea of attunement upon which the domains structuring our methodology (see Sensorimotor dimension: A – the

SENSORIMOTOR DIMENSION	BEHAVIOURAL DIMENSION	EMOTIONAL DIMENSION
A-The bodies	B -The doing	C-The feelings
A1-The space	B1-The doing together	C1-The feeling together
A2-The body of the other	B2-The mutual gaze while doing together	C2-The emotional gaze
A3-The infant's own body	B3-Joint attention	C3-The facial expression
A4-Degree of attention to the motor sanction of the caregiver	B4-Degree of attention to the behavioural sanction of the caregiver	C4-Degree of attention to the emotional sanction of the caregiver

Table 5.1: Description of the features.

bodies; Behavioural dimension: B – the doing; Emotional dimension: C – the feeling) is not aimed at analyzing or considering infants' specific cognitive competences and skills. Instead, following a semiotic perspective, our methodology focuses on their general capacity and willingness to manage the meaning production and recognition through the various phases of the practices in which they are involved. Thus, our methodology aims at analysing and evaluating infants' competence to manage sense-making processes through interaction. In this vein, our methodology is consistent with the methodological criteria and aims of ADOS-2, the goal of its activities being "not to test specific cognitive abilities or other skills, but to present tasks that are sufficiently intriguing so that the child or adult being assessed will want to participate in social interchanges" [Lord, 1999]. As our addressee is an ordinary observer, we have operated with a view to simplification to reduce the heterogeneity and complexity of the searchable signs identified in the diagnostic screening tests. Simplification means that all signs must be summed up in a small number of things to look for that a caregiver can easily evaluate. The hardest part of this work has been removing all of the semiotic technicalities that have been used in order to accomplish that and ending up with something that can be told in a very simple and clear way: if the infant attunes to the caregiver, he/she is essentially a typical-developing infant; if not, the child needs to be monitored, as children who do not tune in to their caregivers during interactions often receive a diagnosis of ASD or other neurodevelopmental disorders later. The main aspect that makes the system very simple is that one only needs to observe the attunement between the infant and the caregiver during their interaction. So, what can be attuned in an interaction? We have identified three main dimensions of attunement: A) the bodies; B) the doing; C) the feelings, a sensorimotor, a behavioral, and an emotional dimension (See Table 5.1).

Sensorimotor dimension: A - The bodies

The four items structuring this first dimension aim to analyze infant-caregiver interaction on its sensorimotor features. Indeed, infants later diagnosed with ASD present a lack of motor control, as they seem not able to coordinate and balance the movements between limbs, trunk, and head [Teitelbaum et al., 1998].

This first dimension presents four signs useful to identify sensorimotor anomalies, which prevent the development of an attuned dynamic of interaction. The four aspects considered are:

Space (A1): Examines the distance between subjects, the movement towards/away from each other, and the typicality of the infant's approach to caregivers or other infants. ASD children may display differences in personal space interactions, such as staying too close or maintaining excessive distance.

Body of the Other (Bodily Attunement - A2): Assesses the extent to which the child adapts their body to caregivers or other children during physical encounters. ASD infants often fail to adapt their posture and movements to align with another person's, appearing rigid, controlled, and inattentive.

Infant's Own Body (A3): Measures the child's overall body posture and movement style, including non-interactive situations. Infants later diagnosed with ASD may exhibit sensorimotor deficiencies, repetitive body movements (stimming), and rigid postures.

Degree of Attention to Motor Sanction of the Caregiver (A4): Focuses on the child's attention and anticipation of caregivers' body movements, especially during actions requiring a specific reaction from the infant. ASD infants may show a reduced ability to react appropriately to others' actions, and mimic behaviors are often absent or reduced.

Behavioural dimension: B - The doing

The second dimension of analysis focuses on the observable behaviors of infants referred as the "doing" dimension, divided into four categories:

Doing Together (B1): Measures the child's spontaneous participation in shared activities, especially in unstructured and unplanned situations. It assesses the child's ability to fluently adapt to changes in a game or task and play a role within it. The evaluation aligns with the core activities and criteria of ADOS.

Mutual Gaze While Doing Together (B2): Examines the frequency and style of the child's eye contact with caregivers or other infants during

shared activities. Reduced or absent eye contact and a lack of attention to others' faces may indicate potential ASD-related concerns.

Joint Attention (B3): Measures the extent to which the child's attention synchronizes with the caregiver's during shared games or tasks. Difficulty in attuning attention, reduced communication, and a lack of engagement with others may be warning signs, as the child may appear isolated and less responsive to the surrounding social context.

Degree of Attention to Behavioral Sanction of the Caregiver (B4): Assesses how well the infant perceives, is aware of and reacts appropriately to context-relevant actions and gestures made by the caregiver during shared activities. This includes the child's understanding of the meaning of actions, games, or tasks and their responsiveness to encouragement and instructions. Children with ASD may exhibit less responsiveness to positive gestures, hindering their social learning and ability to form social bonds.

Emotional dimension: C - The feelings

The third dimension of analysis focuses on emotions in the context of interactions. It examines how these children respond to changes in others' emotions, emphasizing that attunement does not mean sharing the same emotion but involves considering and adjusting to the emotions of others. The dimension is subdivided into four categories:

Feeling Together (C1): Measures how the infant and caregiver adjust their emotional states in response to each other. It looks for signs of the infant becoming happy when the caregiver is happy and the ability to adjust this emotion based on subtle cues from the caregiver. Attunement involves a dynamic and spontaneous adaptation of emotional states between the infant and caregiver.

Emotional Gaze (C2): Evaluates the frequency of eye contact between children and caregivers outside of task-related contexts. Natural and spontaneous eye contact contributes to the overall quality of interaction. A potential concern arises if the child shows disinterest in making eye contact or seems to avoid it, especially when interacting with others.

Facial Expressions (C3): Measures the extent to which the child imitates or reacts spontaneously to caregivers' facial expressions. Unlike the general emotional state assessed in C1, this category focuses on specific facial expressions and how the child responds to the caregiver's emotional cues. An alarming situation may occur if the child remains unaware of the meaning behind the caregiver's facial expressions.

Degree of Attention to Emotional Sanction of the Caregiver (C4): Assesses how attentive the child is to the caregiver's emotional cues and requests. An alarming situation may be indicated if the child shows less interest in the emotional states of others, fails to respond to prompts, or does not exhibit the expected emotional reactions in line with the caregiver's cues.

5.1.3 The tool and the rating procedure

Based on these domains articulating the interactions between infants and caregivers, we built a tool, in which items and rating systems aim at individuating signs of potential concern. Indeed, one of the main problems in ASD screening tests concerns its questions and rating criteria. Thus, in order to bring forth the ecological approach implied by our observational methodology, which emphasizes the prominence of the occurring situation and contextual variabilities of the behaviors, we associated each item of the three dimensions with a rating scale from 1 to 8. In particular, as detailed in the manual using video examples, 1 stands for a very typical interaction (high level of attunement) while 8 stands for a very atypical interaction (low level of attunement). Depending on the severity of the condition and the number of anomalies detected, we divide each field into a range of possible concerns – where 1–2 means no concern, 3–4 means light concern, 5–6 means mild concern, and 7–8 means severe concern. This approach has the major advantage of sterilizing the observers' emotions and points of view. Indeed, the observers do not make any diagnosis: they simply evaluate a behavior. The observers also have the possibility not to rate one (or more) of the features if the home video they are watching is not explicit on that particular aspect (too short, unclear, etc.) or if they feel unsure or incapable of doing so. However, in the first case, they are invited to insert a "not readable" value in the sheet, while, in the second case, they are encouraged to get acquainted with the videos and with the system, trying to read the manual again and train themselves with new material. It will be the system that will later tell whether the child recorded in that particular interaction is behaving typically or not (see Results). They were home videos of i) children who were later diagnosed with ASD when they were 9–18 months old and ii) typically developing children who were used as a control group.

We recruited a total of 32 children with a diagnosis of ASD (10 individuals, 2 females) and typical development (TD) children (22 individuals, 10 females). At the recruitment stage, the children were between 18

months and 11 years old, while all the collected videos show them interacting with caregivers during their 9–18 months period. All participants were recruited after the project dissemination through social media and word of mouth. ASD and TD children were sex-matched ($\chi 2 = 1.9$, $p - value = 0.16$). A total of 67 home videos were collected (28 for the ASD group; 39 for the TD group); for each child, we collected 1.86 ± 1.3 (mean ± standard deviation) videos for the ASD group and 1.77 ± 1.0 for the TD group. Each video was made by parents or caregivers when the children were aged 9–18 months. They were recorded using smartphone cameras during their daily activities to cover the aspects of the three dimensions of interaction.

Measurement of the inter-observer agreement

We measured the inter-observer agreement of the features using the linearly weighted kappa statistic [Altman, 1999]. We adopted the classes of interpretation of the kappa statistic proposed by Landis and Koch [Landis and Koch, 1977] for descriptive purposes. Accordingly, a kappa value below 0 indicates poor agreement, between 0 and 0.20 a slight agreement, between 0.21 and 0.40 a fair agreement, between 0.41 and 0.60 a moderate agreement, between 0.61 and 0.80 a substantial agreement and between 0.81 and 1 an almost perfect agreement (see Table 5.2).

ASD vs. TD classification through an XAI approach

We built a machine learning model for ASD/TD children classification based on XGBoost. We fed the machine learning model by all attunement features and the sex of the children. We trained and tested the ASD vs. TD binary classification task using nested CV [Mueller and Guido, 2017]. In particular, in this study we performed a nested CV loop with a stratified child-based group data splitting scheme to examine the unbiased generalization performance of the trained model and, at the same time, perform hyperparameters optimization (See Fig. 5.1) [Mueller and Guido, 2017].

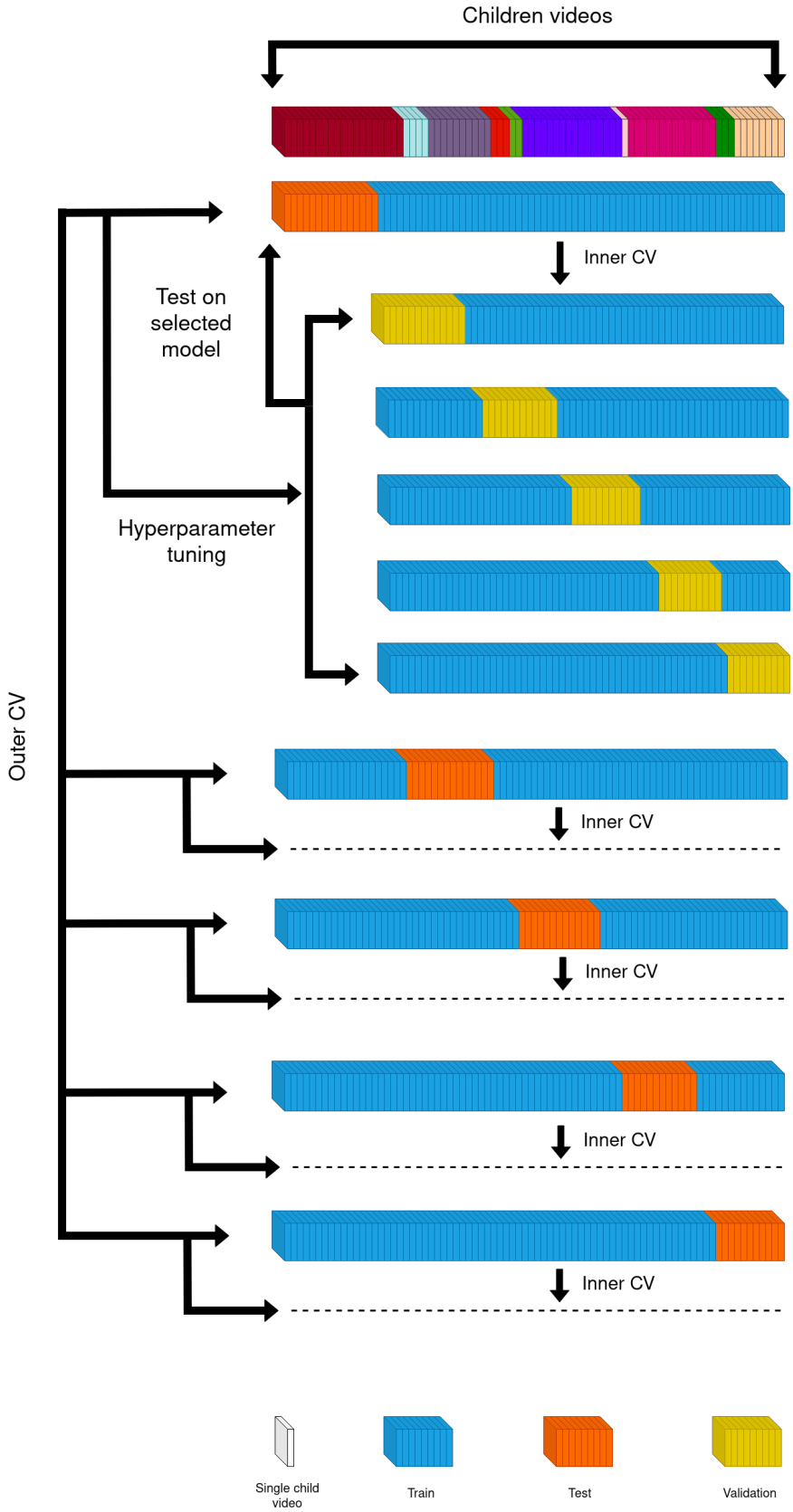


Figure 5.1: Five-fold nested CV with a stratified child-based group data splitting scheme.

In particular, this strategy involves nesting two k-fold CV loops where the inner loop is used for optimizing model hyperparameters, and the outer loop gives an unbiased estimate of the performance of the best model. The stratified child-based group data splitting scheme in the inner and outer CV prevents data leakage, i.e., videos belonging to the same child may be included, at the same time, in the training/validation/test sets. Briefly, we start 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 split into k inner folds ($k-1$ for training and the k_{th} for validation). Specifically, we used a 5-fold nested CV because it offers a favorable bias-trade-off [Hastie, 2013, Lemm et al., 2011]. The models' hyperparameters are chosen from the hyperparameter space through a grid search based on the average performance of the model over the validation sets of the inner folds. In particular, we varied the gamma hyperparameter of the XGBoost in the set {0, 1, 2}, maximum depth in {4, 5, 6, 7, 8}, minimum child weight in {1, 2, 3, 4, 5}, and maximum delta step in {1, 3, 5, 7}. Once the best combination of hyperparameters that maximized the area under the Receiver Operating Characteristic (ROC) curve in the validation sets of the inner CV 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. This procedure is repeated for each fold of the outer CV. Before training each XGBoost classifier in the inner and outer CV, we first applied feature imputation, i.e., we replaced missing features with the average value of that feature in the set. Secondly, the set was standardized, i.e., each feature was rescaled to have zero mean and unit variance. For each iteration of the inner and outer CVs, these transformations were applied to the training, test, and validation sets using Python scikit-learn transformers, thus not using test data in any way during the learning process, – preventing any form of peeking [Diciotti et al., 2013]. Performance was quantified in terms of the AUC of the ROC curve in the test sets of the outer CV. The point of the ROC curve with minimum distance from the ideal observer's performance (0,1) was also computed for both observers. Since the performance may vary depending on how the data are split in each fold of the CV, we repeated the nested CV procedure ten times. We took the average and standard deviation of the results from all repetitions to get a final model assessment score. Since we were interested in explaining the model predictions, we adopted representative SHAP values [Scheda and Diciotti, 2022]. Accordingly, for each model of the outer CV, SHAP values were computed to produce an average and standard deviation of the feature importance explanation for the final model. The training, validation, and test of the XAI

models were carried out using a custom code in Python language (v. 3.8.8) using the following modules: matplotlib v.3.4.1, numpy v.1.21.3, pandas v.1.2.3, scikit-learn v.1.1.dev0 [Pedregosa et al., 2011], seaborn v.0.11.2, and xgboost v.1.4.2. The total computation time for the training, validation, and test was about 40 min on all cores of a Linux workstation equipped with a 4-core (4 threads) Intel i7-7500U CPU and 8 GB RAM.

Feature	Observer #1		Observer #2		Weighted Kappa Statistic
	TD group	ASD group	TD group	ASD group	
A1	1.54 ± 1.2 (10.2%)	4.42 ± 2.3 (25.0%)	1.63 ± 1.2 (10.3%)	3.84 ± 2.6 (32.1%)	0.81
A2	1.54 ± 0.9 (5.1%)	5.64 ± 1.77 (21.4%)	1.63 ± 1.0 (5.1%)	4.91 ± 2.0 (21.4%)	0.85
A3	1.92 ± 1.4 (2.6%)	6.36 ± 1.7 (0.0%)	1.95 ± 1.4 (2.6%)	5.82 ± 2.1 (0.0%)	0.89
A4	2.03 ± 1.4 (12.8%)	6.34 ± 1.8 (17.9%)	2.06 ± 1.6 (12.8%)	5.23 ± 2.2 (21.4%)	0.81
B1	1.54 ± 1.1 (5.1%)	4.96 ± 2.7 (14.3%)	1.54 ± 1.2 (5.1%)	5.21 ± 2.6 (14.3%)	0.95
B2	1.65 ± 1.3 (5.1%)	5.48 ± 2.36 (3.6%)	1.68 ± 1.5 (5.1%)	5.61 ± 2.4 (0.0%)	0.90
B3	1.76 ± 1.4 (5.1%)	5.20 ± 2.6 (10.7%)	1.68 ± 1.4 (5.1%)	5.10 ± 2.4 (10.7%)	0.94
B4	1.63 ± 1.4 (5.1%)	5.11 ± 2.6 (3.6%)	1.66 ± 1.5 (2.6%)	5.28 ± 2.6 (10.7%)	0.94
C1	1.51 ± 1.0 (10.3%)	6.12 ± 2.1 (7.1%)	1.34 ± 0.95 (10.3%)	6.08 ± 2.3 (7.1%)	0.91
C2	1.65 ± 1.6 (5.1%)	5.88 ± 2.3 (10.7%)	1.65 ± 1.6 (5.1%)	5.88 ± 2.2 (7.1%)	0.98
C3	1.53 ± 1.3 (7.7%)	6.00 ± 2.2 (7.1%)	1.50 ± 1.3 (12.8%)	6.23 ± 2.2 (7.1%)	0.94
C4	1.63 ± 1.3 (10.3%)	6.24 ± 2.2 (10.7%)	1.53 ± 1.1 (12.8%)	6.00 ± 2.2 (10.7%)	0.93

Table 5.2: Descriptive statistics of the features measured by both observers [$mean \pm SD^a(\%unassigned)$]. [a] SD: standard deviation.

5.1.4 Results

The model trained with features extracted by observers #1 and #2 reached an AUC of 0.938 and 0.914, respectively – thus indicating an excellent performance of the classification models (see Figure 5.2). The point of the ROC curve with minimum distance from the ideal observer's performance (0,1) corresponds to a sensitivity = 0.89, specificity = 0.86 for observer #1 and sensitivity = 0.85, specificity = 0.86 for observer #2.

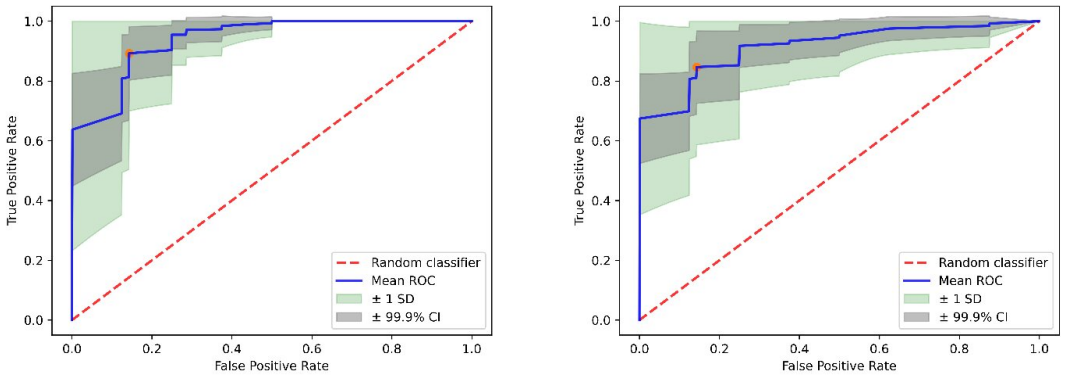


Figure 5.2: ROC curve, in the test set of the outer CV, along with one standard deviation and 99.9% confidence interval curves of the models trained with features extracted by (a) observer #1 (AUC =0.938 $\pm$ 0.015), and (b) observer #2 (AUC = 0.914 $\pm$ 0.021). The point of the ROC curve with minimum distance from the performance of the ideal observer (0,1) is indicated with a red point and corresponds to a sensitivity = 0.89, specificity = 0.86 for observer #1, and sensitivity = 0.85, specificity = 0.86 for observer #2.

In the global feature plot, each feature's global importance is assumed as the mean absolute SHAP value for that feature over all the given samples expressing the average impact on model output magnitude. In Figure 5.4, we showed the ranking of the feature importance, i.e., the SHAP strength (the absolute value of the SHAP values) for both observers. As an example, in Figure 5.3, we plotted the beeswarm plots for the models trained in the first repetition. These plots are designed to show an information-dense summary of how the principal features in the dataset impact the model's output. For each video, the given explanation is represented by a single dot on every single feature. The SHAP value of that feature defines the x position of the dot, and dots "pile-up" along each feature row to show density. Colors are used to display the original value of a feature. In other words, from these plots, we can see the value of SHAP for each sample and each feature. The SHAP value of a given sample (a dot point in the graph) takes on a different color (from blue to pink) whether it has a high or low value, while it takes a different position in the graph (from the base value of SHAP to the right or left), based on its impact on the decision of the model. Therefore, if we observe a SHAP value colored pink and located to the right, the feature has considerably affected the model's decision. In particular, for observer #1, high values of feature A2

move the model's output toward the ASD group, while for observer #2, a low value of feature C1 moves the model's output toward the TD group. For both observers, the model gave considerable importance to the sex of the children, and if the sex is male (coded as '1') the model's output moves toward the ASD group.

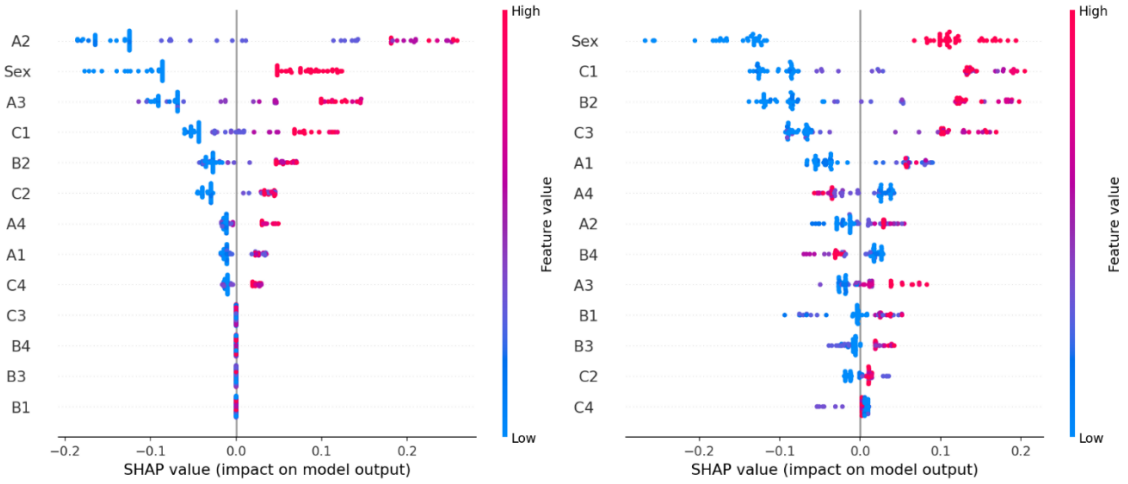


Figure 5.3: The SHAP beeswarm plot is shown for the models trained in the first repetition for (a) observer #1 and (b) observer #2 (see Table 5.1 for the description of the features). High feature values for sex are coded for "male".

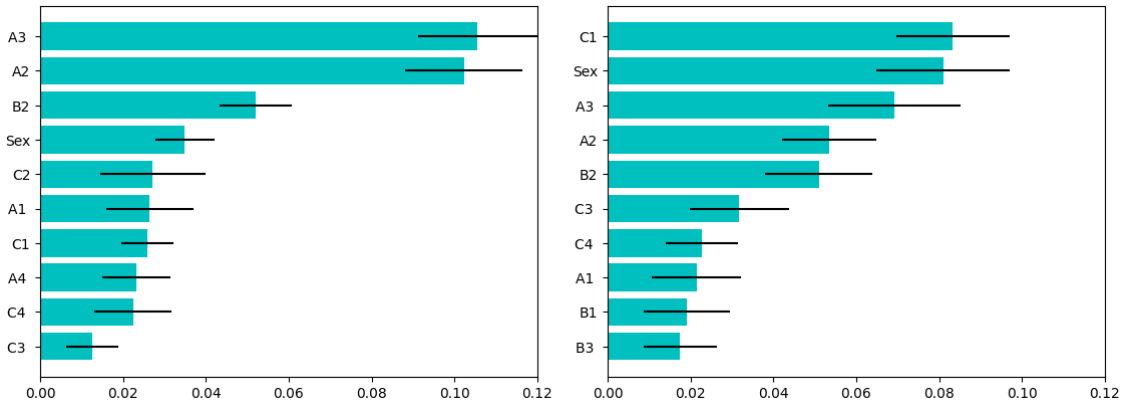


Figure 5.4: Representative SHAP values (average and standard deviation of the absolute values among the ten repetitions) for (a) observers #1 and (b) #2 (see Table 5.2 for the description of the features). The first ten features in the ranking have been shown only.

5.1.5 Discussion

In this study, we predicted ASD classification through interaction features extracted according to our observational methodology by experienced observers watching children's videos during their daily activities [Paolucci, 2021]. We found that the inter-observer agreement between the two observers was almost perfect. Then, we trained a machine learning model with features extracted by observers #1 and #2, reaching excellent AUC in discrimination ASD vs. TD children (Fig. 5.2). Then, we gave XAI interpretability through representative SHAP values to evaluate which of the newly proposed features were effective in classifying individuals with ASD compared to healthy individuals. For observer #1, the features A3 and A2 both concern "The bodies" and the sensorimotor dimension, which were the most crucial in classifying ASD vs. TD subjects, while observer #2 individuated feature A3 was the third most salient. Maybe this "supremacy" and heuristics of the body – if compared to the doing and the emotions – to reveal ASD in prelinguistic infants during secondary intersubjectivity is worth further reflection. As previously stated (see Introduction), the classical cognitive sciences thought of cognition as a representational mental faculty, in which perception used to be conceived as the input and action as the output. Indeed, according to this model, cognition was grounded on largely disembodied mental computations, in which the body certainly played a secondary role, albeit a secondary one, compared to the centrality of the mind and its meta-representational abilities. The classical way of thinking, diagnosing, and understanding autism spectrum disorders has been strongly influenced by this framework: ASD has been considered a Theory of Mind disorder. In other words, ASD was regarded as a disturbance of the metarepresentation capacity of the constitutive type of mind that enables us to make cognitive sense of the actions of others. Indeed, classical and impactful studies, such as those by Baron-Cohen [Baron-Cohen et al., 1985, Baron-Cohen, 1995, Baron-Cohen, 2000] have linked the inability of ASD children to pass meta transactional tests of false belief well into adulthood (and beyond) precisely to a deficit in their social cognition skills. However, as cognitive sciences gradually abandoned the so-called model which identified cognition with the processing of mental representations while increasingly considering perception and action as constituent parts of cognition, the centrality of the theory of mind for ASD began to weaken. On the other hand, the problematic nature of this centrality fits perfectly with the following finding: typical-developing children themselves do not pass theory-of-mind tests until about age 3–4 years, whereas we are capable of discriminating be-

tween population groups with ASD and control groups much earlier with excellent performance. Thus, the AI analysis presented here shows that the body-related signs are the most effective in discriminating between ASD and the control group. In particular, attunement to the other's body (A2) and the signs related to owning one's body (A3) seem to be extremely revealing of a possible future typical autism spectrum disorder, even in very young children, and also in contrast to the other signs from the other two dimensions. It should be noted that the most important features are not exactly the same between the two observers. However, the XAI model clearly points us in the right direction: aiming for a simplification of the system that preserves its performance and robustness while at the same time indicating which of the core signs may better reveal a potential ASD case than others. This was made possible by the SHAP approach which can compute the importance of each feature. However, this study has several limitations. The first concern relates to the observer. Two competent observers have led this pilot study, but the ideal aim of the methodology is to be efficient and apt for non-competent observers. The second concern concerns the sample size, consisting of relatively few children. Thirdly, a disproportion between male and female children is denoted by a small number of females (i.e., 12 out of 32 total children). This condition is frequently encountered in the study of ASD. It is partly due to genetic reasons and partly because the clinical scales used to date are mainly made on male subjects, thus resulting in females being underdiagnosed. Therefore, the results might be more calibrated to males than both sexes. Eventually, the most rated features are not completely consistent with one another. The common features statistically evaluated as salient by both observers are A2 and A3, respectively relating to infants' capacity to adapt his/her body to caregivers or other children during physical encounters (A2) and overall body posture style and movement style, even in non-interactive situations (A3). The prominence of bodily features seems to confirm Teitelbaum and colleagues' study [Teitelbaum et al., 1998], in which, through the analyses of recorded home videos, they identified anomalies in sensorimotor and bodily movements for ASD disorders since the very first months of life. However, despite the inconsistency in the overall balance between the most rated features, the results generated by our AI system prove to be highly sensitive and specific for individuating potential signals. More observers will perform future analyses to determine if it will be possible to individuate a more consistent number of features rated statistically, possibly to create an even simpler yet sensible and specific set of items. In conclusion, our results suggest that examining videos of children engaged in their daily activities through an explainable machine learning

algorithm allows the validation of the effectiveness of some of the constituent features of the proposed scale in classifying ASDs.

5.2 Explainability in dementia transition diagnosis

The study reported in this section refers to the published journal paper entitled *"Fractal dimension of the cortical gray matter outweighs other brain MRI features as a predictor of transition to dementia in patients with mild cognitive impairment and leukoaraiosis"*, Marzi C, Scheda R, Salvadori E, Giorgio A, De Stefano N, Poggesi A, Inzitari D, Pantoni L, Mascalchi M, Diciotti S.

In this work, representative SHAP values were used for evaluating the most important features in predicting the transition to dementia. Indeed, this longitudinal study sought to investigate MRI characteristics that could potentially anticipate the shift to dementia in mild cognitive impairment (MCI) patients with T2 hyperintensities in the cerebral white matter (WM), known as leukoaraiosis. Over two years, 64 participants with MCI and moderate to severe leukoaraiosis underwent baseline MRI examinations and annual neuropsychological assessments. Demographic, neuropsychological, and MRI features at baseline were assessed as potential predictors of clinical transition. These encompassed visually assessed MRI traits like lacune count, microbleeds, and dilated perivascular spaces, as well as quantitative MRI parameters such as volumes of cortical grey matter (GM), hippocampus, T2 hyperintensities, and diffusion indices of cerebral WM. Additionally, we explored advanced quantitative features, including the fractal dimension of cortical GM and WM, derived from 3D-T1 weighted images, serving as an indicator of tissue structural complexity. Representative SHAP values brought out Cortical GM FD as the most influential predictive feature of the transition. Moreover, representative SHAP values of the combined quantitative neuroimaging features demonstrated superior performance compared to visually assessed MRI features in forecasting the transition to dementia.

5.2.1 Introduction

Mild cognitive impairment is a condition marked by inconsistent cognitive function impairment that does not disrupt daily activities [Gauthier et al., 2006]. Over fifty percent of individuals with MCI advance to dementia within the subsequent five years [Gauthier et al., 2006]. Given the overlap of vascular and neurodegenerative diseases in the elderly population and their potential involvement in both MCI and dementia [Jellinger, 2013], discerning the specific contributions of each to the transition to de-

mentia can pose a challenging task. Alterations in the subcortical white matter (WM) of the brain, characterized by areas of reduced density on computed tomography or hyperintensities on T2-weighted MR images, known as leukoaraiosis, are linked to modifications in the diffusion of water protons. This phenomenon is observed in T2-weighted hyperintense and seemingly normal-appearing WM [O'Sullivan, 2008]. These WM changes are prevalent among elderly individuals whose cognitive abilities range from standard to mild cognitive impairment (MCI) and dementia [O'Sullivan, 2008, Fazekas et al., 1987, Golomb et al., 1995, Inzitari et al., 2009]. Leukoaraiosis, along with lacunes and microbleeds, is a marker of small vessel disease (SVD) [Jokinen et al., 2015, Jokinen et al., 2020, Lambert et al., 2016, Williams et al., 2017, Williams et al., 2019, Zeestraten et al., 2017] but, overall, it is a nonspecific finding being observed in elderly subjects with preserved cognition and patients with Alzheimer's disease (AD) [Fazekas et al., 1987, Golomb et al., 1995, Bracco et al., 2005, O'Sullivan, 2008, Bilello et al., 2015]. The Vascular MCI (VMCI) Tuscany study was aimed to identify clinical, neuroimaging, and biological markers predictive of transition to dementia in patients with MCI and leukoaraiosis [Poggesi et al., 2012]. Visually assessed MRI features of brain damage in the VMCI Tuscany included the number of lacunes, microbleeds [Valenti et al., 2016], and dilated perivascular spaces [Mascalchi et al., 2014]. Quantitative MRI assessment included volumes of the entire cortical gray matter (GM), hippocampus, and T2 hyperintense WM [Giorgio et al., 2019], and diffusion properties of the T2 hyperintense and normal-appearing WM [Mascalchi et al., 2014, Ciulli et al., 2016]. We also considered advanced quantitative features such as the fractal dimension (FD) of the cortical GM and WM [Pantoni et al., 2019a] – indices of tissue structural complexity extracted from 3D-T1 weighted images [Marzi et al., 2020]. In this study, our objective was to assess the capabilities of demographic, neuropsychological assessments, visually assessed MRI features, and quantitative MRI features in predicting the transition to dementia over a 2-year span.

5.2.2 Materials and Methods

Feature	Patients without Transition (N=46)	Patients with Transition (N=18)
Demographic		
Age (years)	73.96 ± 6.67 [61.12, 89.03]	76.34 ± 6.693 [59.80, 84.09]
Sex	22 female and 24 male patients	8 female and 10 male patients
Education	8.17 ± 4.25 [3, 18]	7.44 ± 4.30 [2, 18]
Neuropsychological Tests		
MoCA	21.23 ± 4.62 [11.95, 29.29]	18.93 ± 3.95 [13.10, 25.24]
ROC-F Immediate Copy	23.68 ± 7.21 [5.59, 35.58]	21.27 ± 10.61 [4, 36]
SDMT	39.18 ± 10.03 [22.02, 59.94]	31.18 ± 5.38 [24.67, 43.49]
Stroop	33.44 ± 23.81 [-3.45, 114.57]	51.59 ± 36.02 [8.83, 155.09]
TMT-A	61.47 ± 47.97 [3.77, 202.2]	64.47 ± 43.15 [8.42, 152.92]
VS	32.84 ± 8.61 [14.3, 50.17]	29.08 ± 7.78 [15.41, 41.27]
Visually Assessed MRI Features		
Lacunar Infarcts	2.02 ± 0.80 [1, 3]	2.28 ± 0.83 [1, 3]
Cerebral Microbleeds	0.91 ± 2.57 [0, 15]	2.24 ± 5.77 [0, 18]
EPVS Basal Ganglia	1.67 ± 0.82 [0, 4]	1.83 ± 0.62 [1, 3]
EPVS Centrum Semiovale	1.89 ± 0.77 [1, 3]	1.44 ± 0.70 [1, 3]
Quantitative MRI Features		
WM Lesion Load	0.07 ± 0.04 [0.01, 0.20]	0.09 ± 0.05 [0.02, 0.20]
WM Volume	0.15 ± 0.01 [0.12, 0.17]	0.14 ± 0.01 [0.13, 0.16]
GM Volume	0.12 ± 0.01 [0.10, 0.14]	0.11 ± 0.01 [0.10, 0.12]
Hippocampal Volume	0.0023 ± 0.0005 [0.0010, 0.0031]	0.0020 ± 0.0002 [0.0020, 0.0024]
WM FD	2.45 ± 0.04 [2.35, 2.51]	2.43 ± 0.04 [2.36, 2.49]
GM FD	2.34 ± 0.02 [2.30, 2.38]	2.33 ± 0.02 [2.27, 2.36]
Median FA	0.37 ± 0.02 [0.33, 0.41]	0.36 ± 0.02 [0.32, 0.40]
Median MD	0.82 ± 0.05 [0.7, 0.9]	0.82 ± 0.04 [0.8, 0.9]

Table 5.3: Descriptive statistics of demographic, neuropsychological, visually assessed MRI, and quantitative MRI features for patients with and without a 2-year transition to dementia.

In this study the dataset was composed by a cohort of 64 patients with MCI and leukoaraiosis as part of the VMCI Tuscany study. The participants were selected from a single center, and they underwent baseline MRI and annual neuropsychological testing over a period of two years. After two years, 18 (28.1%) participants had converted from MCI to dementia. As described in Table 5.3, the features considered for this study can be divided in four categories:

Demographic features : Age, Sex, Education.

Neuropsycological tests : Each patient underwent a comprehensive neuropsychological evaluation developed for patients with SVD and MCI [Salvadori et al., 2016], including both global cognitive functioning test (i.e., Montreal Cognitive Assessment (MoCA) [Nasreddine et al., 2005, Conti et al., 2015] and second-level tests covering different cognitive domains (i.e., Visual Search (VS) [Sala et al., 1992], Symbol Digit

Modalities Test (SDMT) [Nocentini et al., 2006], Trail Making Test (TMT), Part A , Color Word Stroop Test (Stroop) [Caffarra et al., 2002], and immediate copy of the Rey-Osterrieth Complex Figure (ROC-F)).

Visually Assessed MRI features : An experienced observer visually assessed the number of lacunes (fluid-filled cavities in the brain tissue), cerebral microbleeds (tiny areas of bleeding in the brain), and enlarged perivascular spaces (EPVS) (abnormal expansion of narrow fluid-filled channels surrounding blood vessels in the brain).

Quantitative MRI features : The WM lesion load was calculated as the total lesions' volume normalized by the individual cerebral WM volume. The FreeSurfer image analysis suite v. 5.3 (`http://surfer.nmr.mgh.harvard.edu/`) performed cortical reconstruction and volumetric segmentation of the WM, cortical GM, and hippocampus on T1-weighted images [Fischl, 2012]. We separately computed the hippocampus, WM, and cortical GM volumes in the left and right hemispheres. Then we considered the average value of the volume of each structure in the left and right hemispheres and subsequently normalized them to estimated eTIV. The fractal analysis was carried out using the fractalbrain toolkit version 1.1 [Marzi, 2023]. In this study, we examined the fractal properties of WM and cortical GM by calculating and averaging the FD from the left and right hemispheres of each structure. Finally, fractional anisotropy (FA) and mean diffusivity (MD) were computed: FA describes the degree of anisotropy of water molecules, and MD provides a measure of the directions of diffusion of water molecules. Figure 5.5 shows the extraction procedure of the brain MRI features considered in the present investigation and partially described in detail previously [Pantoni et al., 2019b].

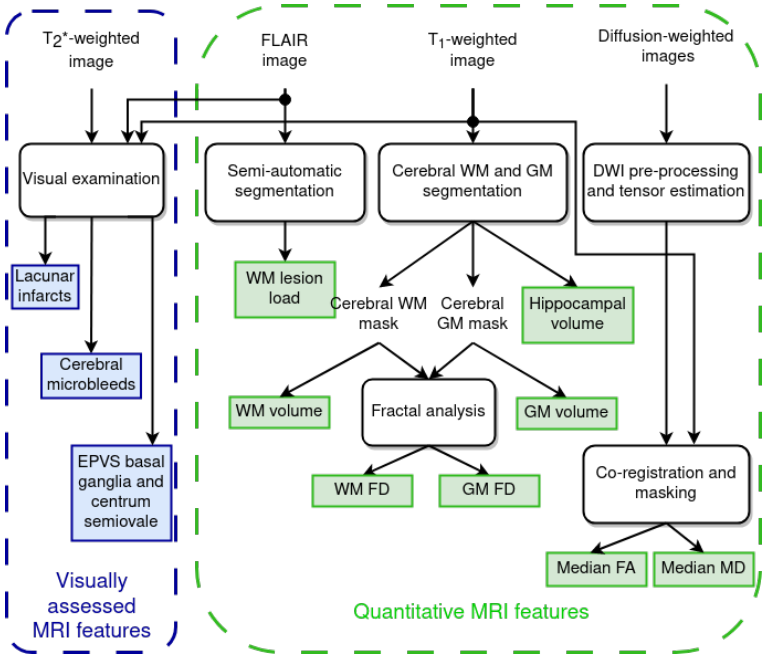


Figure 5.5: Schematic representation of the MRI features extraction for predicting the transition to dementia.

Machine learning system

To forecast the transition to dementia, we used an explainable machine learning framework fed by baseline demographical, neuropsychological, visual, and quantitative MRI features. During the training phase, missing values in the data were imputed by replacing them with the average value of the corresponding feature. Additionally, standardization was performed by rescaling each feature to have a mean of zero and a variance of one. These imputation and standardization techniques were exclusively learned during the training phase and subsequently applied in the validation and testing phases, leading to an unbiased generalization performance. The explainable ML framework was trained, validated, and tested through a repeated stratified nested validation procedure (Figure 5.6).

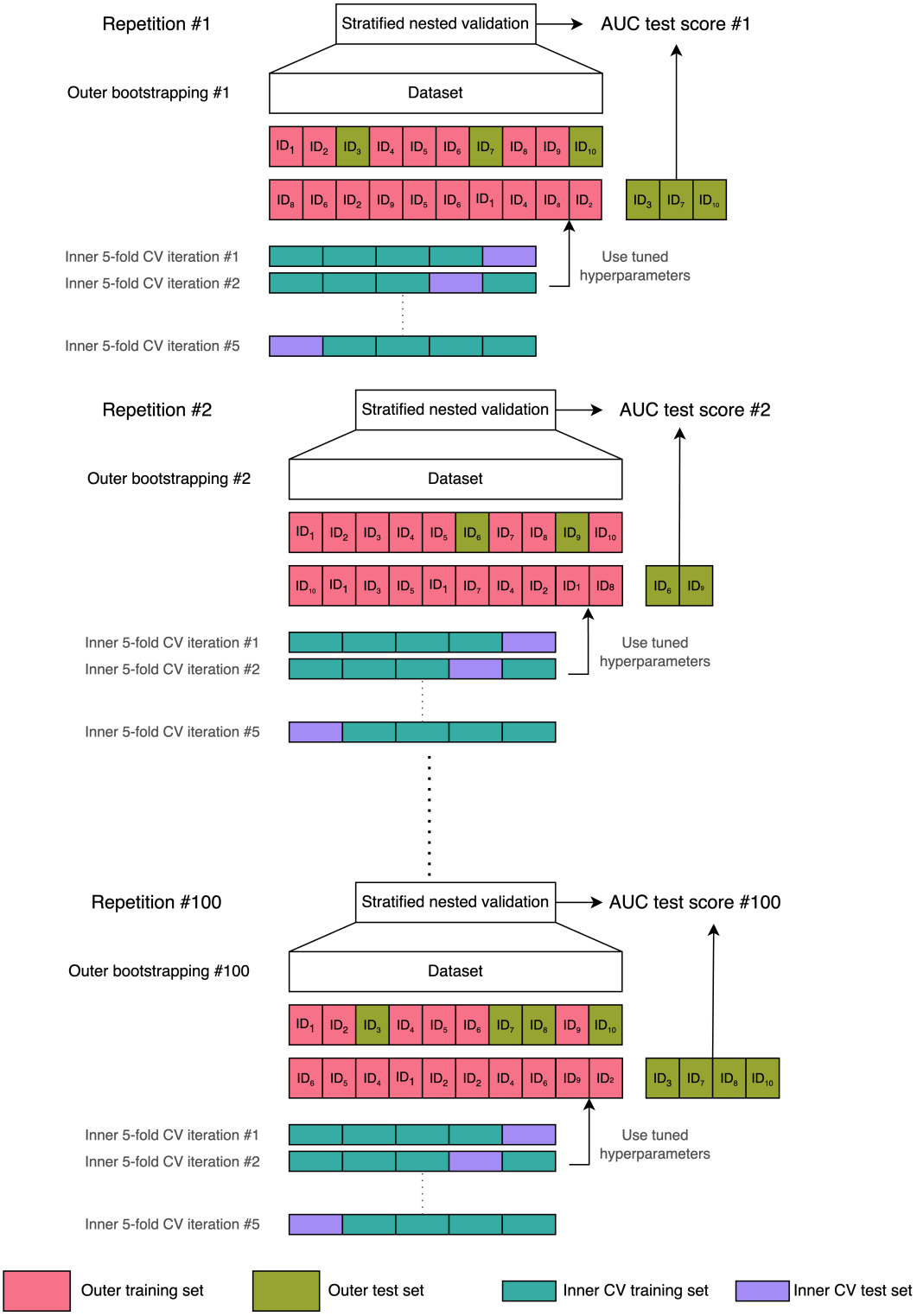


Figure 5.6: Machine learning validation scheme: 100-times repeated stratified nested validation procedure. In the figure, we chose a dataset comprising ten samples to illustrate the bootstrap resampling procedure with comprehensive replacement. When applying bootstrap resampling to our actual dataset, which contains 64 subjects, we obtain an outer training set consisting of 64 instances (some of which are repeated) and an outer test set that comprises the unique instances not included in the training set, referred to as the out-of-bag samples. The outer training set was then used for an inner subject-level 5-fold CV for hyperparameter optimization.

We chose a bootstrap resampling for the outer split and a 5-fold CV for the inner loop. We selected a number of the folds equal to 5 because it offers a favorable bias-variance trade-off [Hastie, 2013]. In detail, for each repetition of the bootstrap resampling, the entire dataset was divided into an outer training set by sampling, with replacement, the instances contained in the original dataset. The outer test set included the unique instances not selected for the training set, i.e., the out-of-bag samples. The outer training set was then used for an inner subject-level 5-fold CV for hyperparameters optimization. The subject-level splitting ensures that the repetitions present in the outer training set are either in the inner training set or in the inner validation set, preventing data leakage [Yagis et al., 2021]. Once the combination of hyperparameters values that minimized the out-of-sample prediction error [Hastie, 2013] has been found in the inner CV, the model with that combination of hyperparameters' values is re-trained on the outer training set and tested on the unseen outer test set, thus preventing any form of peeking effect [Diciotti et al., 2013]. The stratified sampling ensured that samples possessing a particular characteristic, i.e., the transition to dementia, were selected in the same proportion in the training, validation, and test sets as they existed in the entire dataset. The stratified nested validation was repeated 100 times with different bootstrap data splitting to attenuate the dependencies of the model from the training data, along with reducing performance estimation variance while maintaining a minimal bias [Molinaro et al., 2005, Kim, 2009]. The explainable ML framework utilized in this study employed an XGboost model. The model's hyperparameters were selected through a random search within the inner CV process. The hyperparameter space was defined as follows: the minimum loss reduction required for further partitioning a leaf node of the tree $g \in \{0.6, 0.7, 0.8\}$, the subsample ratio of columns used when constructing each tree *colsample_bytree* $\in \{0.25, 0.5, 0.75, 1\}$, the maximum depth of a tree *max_depth* $\in \{2, 3, 4\}$, the minimum number of instances required in each node *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 before growing trees *subsample* $\in \{0.1, 0.2, 0.4\}$.

For each repetition of the stratified nested validation, the classifier performance was evaluated on the outer test set using the area under the ROC AUC curve. The mean AUC and the 90% confidence interval (CI) were reported as the final performance. To verify whether the performance of our classifier was significantly superior to that of a random guessing classifier [Fawcett, 2006], we compared the AUC values with the value 0.5, i.e., the chance-level performance, through a one-tailed Wilcoxon signed rank

with a significance level of 5%. We built a median ROC curve by considering the coordinates of the ROC curve obtained from the data of the outer test set at each repetition of the stratified nested validation. The optimal operating point on the median ROC curve was identified as the point with the highest Youden's index, denoted as J = sensitivity + specificity – 1 [Youden, 1950]. To obtain the feature contributions, representative SHAP values were computed for the outer test set during each repetition of the repeated nested validation. They were subsequently averaged, in absolute value, across patients [Scheda and Diciotti, 2022]. Therefore, we obtained 100 global SHAP values for each feature and calculated the median over the repetitions as the final global feature importance. The global contribution of the top-ranking predictive feature was compared to the second feature of the ranking through a one-tailed Wilcoxon signed rank with a significance of 0.05. In addition to assessing the individual contributions of each feature towards predicting the transition to dementia, we also averaged the SHAP values over specific feature categories (i.e., the sum of the SHAP values of all features belonging to a category divided by the total number of features in the category). These categories included demographics (Age, Sex, Education), adjusted neuropsychological scores (MoCA, TMT-A, ROC-F immediate copy, SDMT, Stroop, VS), visually assessed MRI features (Lacunar infarcts, Cerebral microbleeds, EPVS basal ganglia, EPVS centrum semiovale), and quantitative MRI features (WM lesion load, GM FD, WM FD, hippocampal volume, GM volume, WM volume, Median MD, Median FA). By grouping the SHAP values according to these feature categories, we comprehensively understood their combined contributions to predicting the transition to dementia.

5.2.3 Results

To forecast the transition to dementia, the mean ROC AUC was 0.69 with a 90% CI of (0.53, 0.85). The AUC value of our classifier was significantly higher than the chance-level performance (one-tailed Wilcoxon signed rank p-value < 0.001). Through ROC curve analysis (Figure 5.7), we identified a specific operating point that maximized the Youden's index, gaining a sensitivity of 0.67 and a specificity of 0.67. The GM FD was the top-ranking predictive feature (Figure 5.8).

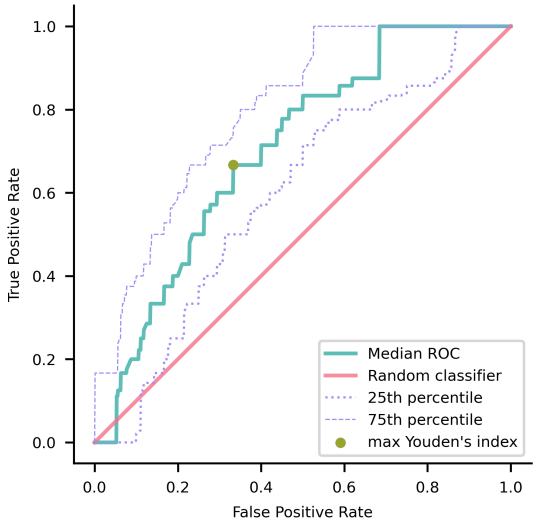


Figure 5.7: Median receiver operating characteristic (ROC) curve of the model trained using nested validation over 100 repetitions. The gold point on the ROC curve corresponds to the coordinates (0.33, 0.67) where the maximal Youden's index is achieved. The red overlay represents the ROC curve of a random classifier, serving as a reference. The dotted and dashed purple curves indicate the 25_{th} and 75_{th} percentiles, respectively.

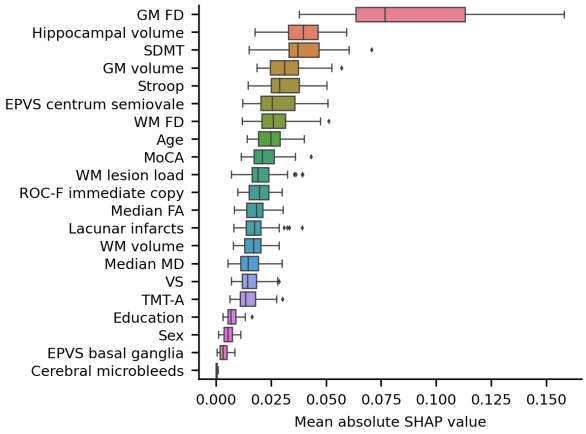


Figure 5.8: A box plot showing each feature's mean absolute SHAP values, sorted in ascending order. The volume and the FD of a specific brain structure are defined as the average values among the left and right hemispheres. Volumes were subsequently normalized to eTIV. EPVS, enlarged perivascular spaces; FD, fractal dimension; GM, gray matter; MoCA, adjusted Montreal Cognitive Assessment score; SDMT, adjusted symbol-digit modality test score; WM, white matter.

The median absolute SHAP value of the GM FD was significantly higher than that of the second-ranking feature, i.e., hippocampal volume (one-tailed Wilcoxon signed rank p-value < 0.001). The main important predictive features were the SDMT score, cortical GM volume, Stroop score, EPVS centrum semiovale, WM FD, age, MoCA score, and WM lesion load. The aggregated quantitative neuroimaging features exhibited superior predictive capabilities compared to visually assessed MRI features (Figure 5.9). Figure 5.10 illustrates the visualization of mean SHAP values corresponding to specific features within individual samples (subjects). These visualizations aim to provide a concise representation of how the dataset's features influence the model's output. A single dot for every feature depicts each subject. A feature's SHAP value determines the dot's horizontal position, and dots accumulate along each feature's row to depict density. Colors are utilized to indicate the original feature values. In essence, these plots enable us to observe the SHAP value for each feature in every sample. In these graphical representations, a dot's color varies (from blue to pink) according to whether the feature value is high or low. Additionally, its position on the graph shifts (from the base SHAP value to the right or left) based on its influence on the model's decision (i.e., its SHAP value). As depicted in Figure 5.10, most dots associated with cortical GM FD are shaded pink and positioned towards the left. This indicates that a lower FD value, showing decreased cortical GM structural complexity, signifi-

cantly impacts the model's decision, guiding it toward the transition to the dementia class.

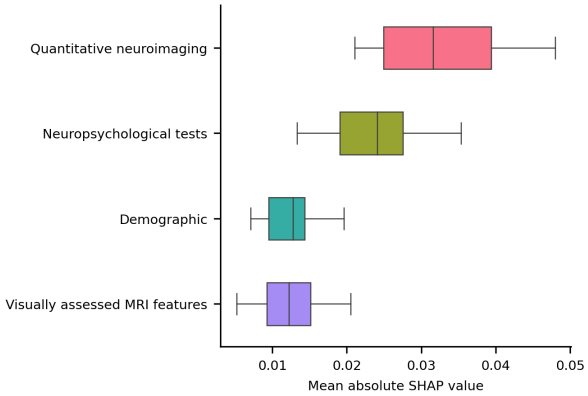


Figure 5.9: A box plot illustrating the averaged absolute SHAP values over each category (i.e., a sum of the SHAP values of all features belonging to a category divided by the total number of features in the category).

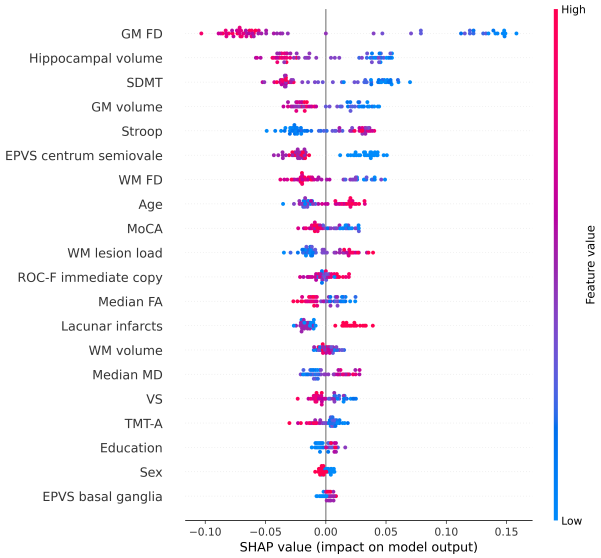


Figure 5.10: Beeswarm summary plot depicting representative SHAP values. Each feature row for each sample (i.e., subject) is represented by a single dot, with the x position determined by the corresponding SHAP value. Dots accumulate along each feature row to indicate density. The color of each dot represents the original value of the feature. A specific brain structure's volume and the FD are defined as the average values among the left and right hemispheres. Volumes were subsequently normalized to eTIV. EPVS, enlarged perivascular spaces; FA, fractal anisotropy; FD, fractal dimension; GM, gray matter; MD, mean diffusivity; MoCA, adjusted Montreal Cognitive Assessment score; ROC-F, adjusted Rey-Osterrieth Complex Figure immediate copy score; SDMT, adjusted symbol-digit modality test score; TMT-A, adjusted trail making test-A score; VS, adjusted visual search score; WM, white matter.

5.2.4 Discussion

Predicting the transition to dementia in patients with MCI is of utmost importance as it could enable the implementation of therapies aimed at slowing or halting the progression of the disease. Interestingly, representative SHAP values have shown that the FD of the cortical GM emerged as the most remarkable best predictor for this transition. Furthermore, FD and volume of the cortical GM exhibited superior predictive performance compared to the WM lesion load, diffusion-derived indices, and FD of the cerebral WM. Notably, when features of the same type were aggregated, quantitative neuroimaging features demonstrated superior predictive capability compared to neuropsychological tests, visually assessed MRI features, and demographic factors. Our findings confirm that corti-

cal GM is closely associated with leukoaraiosis, as demonstrated by previous studies[Lambert et al., 2015, Ye et al., 2015, Heinen et al., 2020]. Moreover, our results highlight the contribution of GM atrophy to transition to dementia in patients with MCI and leukoaraiosis [Jokinen et al., 2012, Jokinen et al., 2020, Bilello et al., 2015, Wu et al., 2019, Fan et al., 2021]. Specifically, it has been observed that cortical atrophy associated with leukoaraiosis exhibits a distinct distribution in the dorsolateral prefrontal, parietal, and posterior-superior temporal cortices, differing from the cortical changes associated with normal aging [Lambert et al., 2015, Ye et al., 2015, Heinen et al., 2020]. Additionally, studies have indicated a correlation between the progression of cortical atrophy and leukoaraiosis over time [Lambert et al., 2016]. Furthermore, hippocampal and medial temporal lobe atrophy have been identified as underlying factors contributing to cognitive deficits in patients with leukoaraiosis [Bastos-Leite et al., 2007, Jokinen et al., 2020, Chen et al., 2021, Fan et al., 2021, Sun et al., 2022] and has been associated with the transition to dementia in these individuals [Jokinen et al., 2012, Jokinen et al., 2020]. The exact nature of cortical changes concerning leukoaraiosis and SVD remains uncertain, as some studies suggest they may be secondary effects of leukoaraiosis/SVD [Bastos-Leite et al., 2007, Jokinen et al., 2020, Chen et al., 2021]. In contrast, others propose a dual pathology involving accompanying AD [Jellinger, 2013, Ye et al., 2015, Wu et al., 2019]. Notably, our study reveals that subtle changes in cortical GM, manifested as decreased FD, better anticipate transitioning from MCI to dementia than overt cortical atrophy. In parallel, it is well-established that "invisible" changes in terms of subtle T2 signal changes [Jokinen et al., 2015] or diffusion properties [Zeestraten et al., 2017, Williams et al., 2019, Egle et al., 2022] can be observed in the normal-appearing WM of patients with leukoaraiosis. These changes are predictive of cognitive decline. In line with these findings, our study suggests that FD of the WM may serve as an additional marker for the subtle structural changes occurring in the WM of patients with leukoaraiosis. The findings of this study further strengthen the evidence that FD provides supplementary information beyond what is offered by other conventional structural features [Free et al., 1996, Im et al., 2006, Sandu et al., 2008b, Sandu et al., 2008a, Sandu et al., 2014a, Sandu et al., 2014b, Sandu et al., 2022, King et al., 2009, King et al., 2010, Madan and Kensinger, 2016, Madan and Kensinger, 2018, Marzi et al., 2018, Marzi et al., 2020, Marzi et al., 2021b, Marzi et al., 2022, Pantoni et al., 2019a, Pani et al., 2022a, Nazlee et al., 2023] and have potential relevant practical and diagnostic implications, particularly regarding the MRI evaluation of the cortical GM. The FD measurement can be derived from standard high-

resolution 3D T1-weighted images commonly included in clinical MRI protocols. This means that FD assessment does not necessitate additional dedicated acquisitions, such as magnetization transfer imaging, which is capable of detecting subtle microstructural changes in the cortical GM in both inherited and sporadic AD [Ginestroni et al., 2009, Mascalchi et al., 2013]. By contrast, nuclear medicine techniques for assessing cortical GM metabolism or amyloid deposits for the differential diagnosis of patients with leukoaraiosis have not been widely implemented [Ye et al., 2015, Altomare et al., 2023]. Therefore, using FD measurement from standard MRI scans may represent a valuable and accessible tool in clinical practice for evaluating cortical GM alterations without requiring additional specialized imaging techniques. We acknowledge several limitations in our study. First, the relatively small sample size and the fact that the study was conducted at a single center may impact the generalizability of our findings. The sample was collected in a highly qualified referral university hospital where patients fulfilling admission criteria were consecutively identified and carefully evaluated before enrollment. Of course, this cannot support the full generalizability of results. Therefore, further validation in independent samples would enhance the robustness and generalizability of the results. Second, considering the whole brain structures rather than regional FD differences does not allow for demonstrating the distributed microstructural or overt changes known to occur in vascular MCI and dementia. Lastly, longitudinal MRI data would be valuable to elucidate the underlying mechanisms better. Unfortunately, such longitudinal data are not available for our study. In conclusion, our study highlights that the transition from MCI to dementia in patients with leukoaraiosis is associated with subtle alterations in the cerebral cortical GM and WM reflected by altered FD. Our findings suggest that the FD changes observed in the cortical GM exhibit a stronger predictive value for future transition than other brain measurements. The FD of the cortical GM emerges as a biomarker potentially more sensitive than other brain measurements for predicting the transition to dementia.

Chapter 6

Swarm learning

This final chapter will focus on the concept of swarm learning. Nodes in a swarm learning system communicate and share information. Swarm learning addresses privacy concerns by allowing data to remain on individual nodes: instead of sharing raw data, nodes exchange model updates or aggregated information, enabling collaborative learning without compromising the privacy of individual data. Different scenarios will be investigated with multiple types of data, analyzing different behaviors and capabilities of this recently-proposed technique.

6.1 Introduction

Several machine learning and deep learning algorithms have been applied to facilitate clinical diagnosis, but such tools often require large clinical datasets for training [Chishti et al., 2020, Al'Aref et al., 2018]. The challenge of acquiring sufficiently large and diverse datasets for training machine learning models in the medical field persists due to various factors. Not only do single-center studies face limitations in sample size due to the intricacies and expenses associated with collecting patient data [Shaikhina and Khovanova, 2017], but the scarcity of data is exacerbated by the rarity of certain diseases [Holzinger, 2018]. In addition to a lack of patient samples, hospitals face stringent privacy policies that restrict or prevent sharing sensitive medical information [Rieke et al., 2020]. This leads hospitals to train machine learning models on very restricted datasets. Consequently, healthcare institutions often find themselves constrained to train machine learning models on poor datasets, impeding the robustness and generalizability of the algorithms. In response to these challenges, emerging techniques such as *Federated learning* (FL) have recently gained popularity. Distributed learning approaches like FL enable model training across multiple institutions without compromising data privacy, offering a promising avenue to address the data scarcity issue and enhance the performance of machine learning models in medical diagnosis and treatment planning. Here we distinguish the different types of learning:

- **Local learning**: data and computation at different and disconnected locations;
- **Central learning**: Data and algorithms are centralized in one cloud-based framework;
- **Federated learning**: data and training being kept private but parameters are orchestrated by a central parameter server;
- **Swarm learning**: Data is kept private, and each node exchanges parameters to train a common model.

Google first introduced Federated learning in 2017 [McMahan et al., 2017] to improve text prediction in mobile keyboards using machine learning models trained by data across multiple devices. This new technology does not require uploading personal data to a central server to train the models, which was a breakthrough in traditional machine learning to address data privacy issues. Swarm learning (SL) is a new kind of distributed learning, which has been recently proposed as a more secure decentralized learning

for deep learning models [Warnat-Herresthal et al., 2021]. The basic concept of swarm learning is that in contrast to FL, swarm learning has no a *central* server that coordinates the merging of the model's parameters. Indeed, each server or center is a *peer*, and at each synchronization step the coordinator server changes. Swarm Learning is based on Blockchain technology, ensuring secure and tracked training between the centers [Warnat-Herresthal et al., 2021].

6.2 Methods

In an SL framework, each center, with its private data, becomes a *swarm node*, which communicates with other swarm nodes (or peers) through the *swarm network*, which is essentially a blockchain overlay on the underlying network connection between the nodes. Here, we introduce the workflow and the processes behind swarm training (see Figure 6.1).

Enrollment

The swarm learning process begins with enrollment, or registration, in the Swarm smart contract by each node. Each node subsequently records its relevant attributes in the contract, such as the uniform resource identifier (URI) from which its own set of other nodes can download trained parameters.

Local model training

Nodes next proceed to train the local copy of the model iteratively over multiple rounds, each called an epoch. During each epoch, every node trains its local model using one or more data batches for a fixed number of iterations. After reaching the number, it exports the parameter values in a file in its local filesystem. Subsequently, it signals other nodes that it is ready for the parameter-sharing step.

Parameter sharing

This step starts once the number of nodes ready for the parameter sharing step reaches a specific minimum threshold value specified during initialization. It begins with electing the *epoch leader*, whose role is to merge the parameters derived after local training on all nodes. This selection is

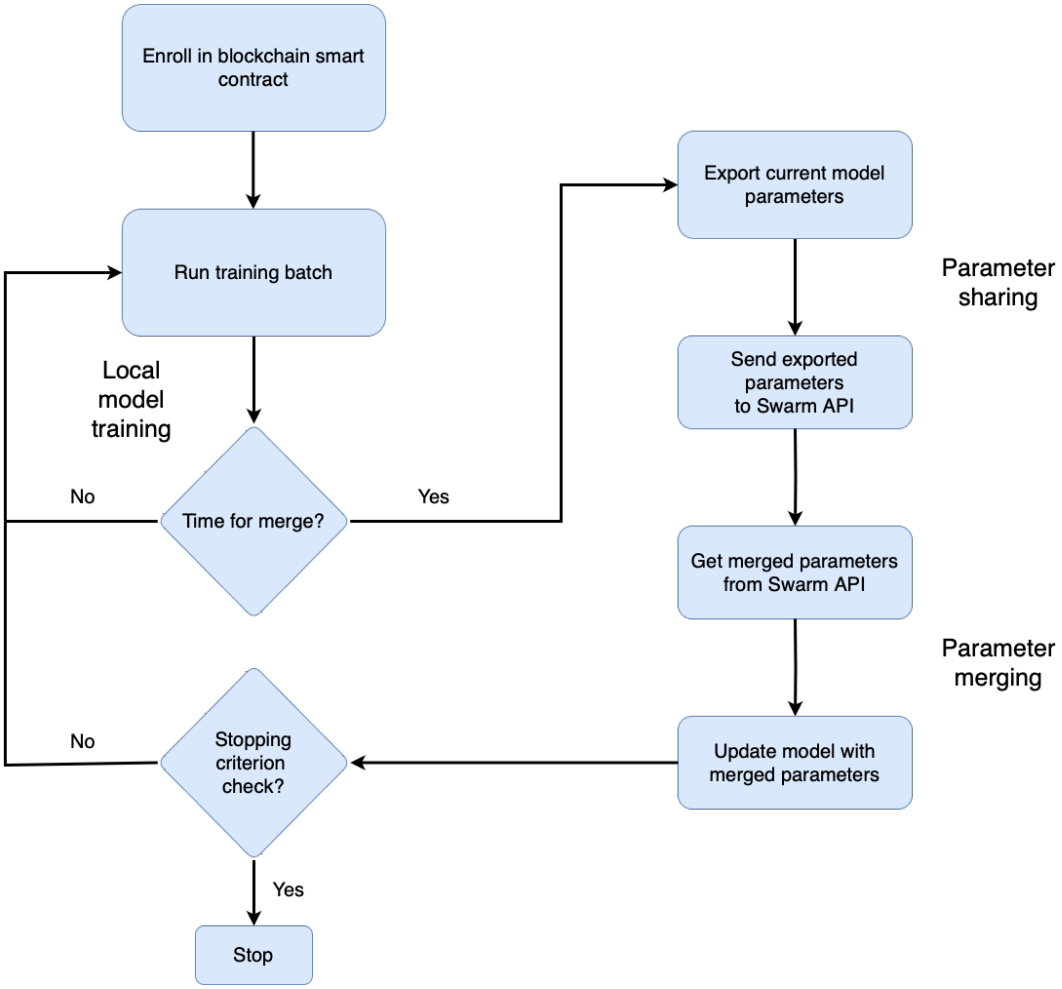


Figure 6.1: Schematic flow chart of swarm training.

rapid and takes place at the culmination of each epoch. Using the predetermined leader election algorithm, one of the nodes emerges as a leader and then downloads the parameter files from each other before the parameter-merging step.

Parameter merging

The leader then merges the parameter files downloaded. The framework supports multiple merge algorithms such as *mean* and *weighted mean*, defined in equation 6.1:

$$P_m = \frac{\sum_{k=1}^{n}(w_k P_k)}{\sum_{k=1}^{n} w_k} \tag{6.1}$$

in which P_M is the $m-$merged parameter, P_k is the parameter from the k_{th} node, w_k is the weight of the k_{th} node, and n is the number of nodes participating in the merge process. Using the merge algorithm chosen, the leader combines the parameter values from all nodes to create a new file with the merged parameters and signals to the other nodes that a new file is available. Each node then downloads the file from the leader and updates its local model with the new set of parameter values.

Stopping criterion

Finally, the nodes evaluate the model with updated parameter values using their local data to calculate various validation metrics. The values obtained from this step are shared. In the official paper presenting swarm learning [Warnat-Herresthal et al., 2021], authors validate local models with the same test set for each node in order to compare local validation metrics. As each node completes this step, it signals to the network that the update and validation step is complete. In the meantime, the leader keeps checking for the updated complete signal from each node. When it discovers that all merge participants have signaled completion, the leader merges the local validation metric numbers to calculate the global metric numbers. The synchronization step is then marked as complete. Afterward, the system's current state is compared against the stopping criterion, and if it is found to be met, the SL process is halted. Otherwise, the steps of local model training, parameter sharing, parameter merging, and stopping criterion check are repeated until the criterion is fulfilled.

6.2.1 Swarm Learning with MNIST dataset

To study the effectiveness of swarm learning, we investigated the swarm learning tool with our servers. We set up three different nodes on the same server (see Figure 6.2): each node had a different portion of the original dataset, trained the local model and merged the parameters with the other nodes, and at the end, each node tested its model with the test set. In order to compare the metrics results, we tested the models of each node with the same test set. The neural network consists of one input layer, one hidden layer, and one output layer. The input layer is densely connected and consists of 512 nodes, a rectified linear unit activation function (ReLU [Agarap, 2018]), and a dropout rate of 20%. The output layer is densely connected and consists of ten nodes and a sigmoid activation function. The model is configured for training with Adam optimization [Kingma and Ba, 2017] and to compute the sparse cross-entropy loss between true

and predicted labels. The model is used for training both the individual nodes and SL. The model is trained over 25 epochs, with a batch size of 64.

MNIST dataset

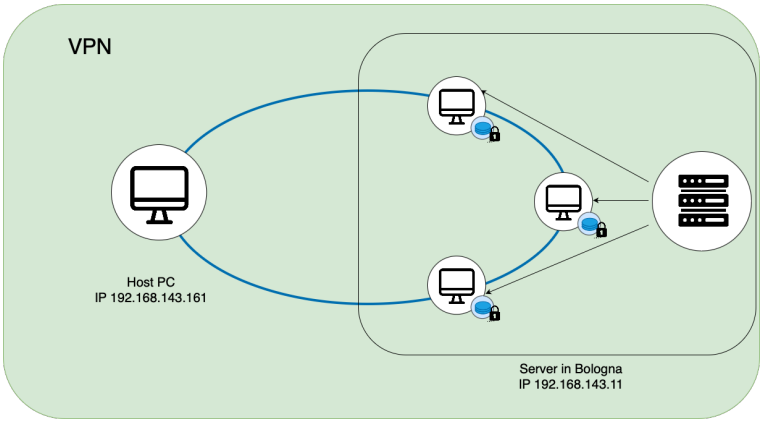


Figure 6.2: Schematic representation of swarm learning set up using the MNIST dataset.

The MNIST dataset [Deng, 2012] is a collection of grayscale images representing handwritten digits, ranging from 0 to 9 (see Figure 6.3).

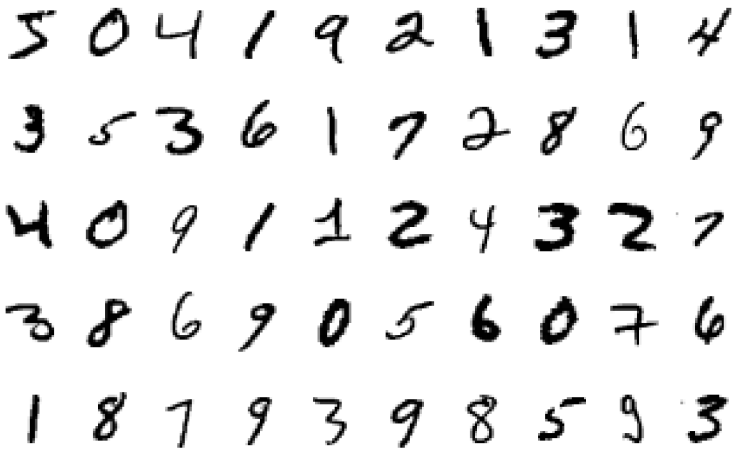


Figure 6.3: Example samples of the MNIST dataset.

Each image in the dataset is a 28x28 pixel square, and it captures the subtle variations in handwriting styles among different individuals. Ini-

tially created for training and testing algorithms for recognizing handwritten characters, the MNIST dataset has become a standard benchmark for evaluating the performance of various machine learning models. The MNIST dataset consists of two principal subsets: a training set and a test set. The training set typically contains 60,000 images, while the test set contains 10,000. The dataset ensures a balanced distribution of digits across both subsets, making it suitable for assessing the generalization ability of classification models.

6.2.2 Swarm Learning with real medical dataset

To study the effectiveness of swarm learning in a real medical scenario, we also investigated the swarm learning tool with three different servers with real medical data. We set up one swarm node in each server: one in Bologna and two servers in Verona (Figure 6.4). Each node had a different portion of the original dataset, trained the local model, and merged the parameters with the other nodes, and at the end, each node tested its model the test set. In order to compare the metrics results, we tested the models of each node with the same test set. The neural network consists of one input layer, three hidden layers, and one output layer. The three hidden layers are densely connected and consist of 16 nodes, a rectified linear unit activation function (ReLU [Agarap, 2018]). The output layer is densely connected and consists of one node and a sigmoid activation function. The model is configured for training with Adam optimization [Kingma and Ba, 2017] and to compute the binary cross-entropy loss between true and predicted labels. The model is used for training both the individual nodes and SL. The model is trained over 25 epochs, with a batch size of 64.

MIMIC III dataset

The MIMIC-III (Medical Information Mart for Intensive Care III) dataset [Johnson et al., 2016] is an extensive and comprehensive electronic health record (EHR) dataset widely used in medical research, particularly in the field of critical care. This dataset comprises deidentified health data from over 40,000 patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts, USA, spanning 11 years. It encompasses diverse clinical information, including vital signs, laboratory results, medications, diagnoses, procedures, and demographics. The dataset reflects the complexity of critical care scenarios, capturing the dynamic nature of patient conditions over time. The dataset is organized into various tables,

each representing different aspects of patient care. The tables are linked through unique patient identifiers, enabling researchers to perform complex queries and analyses across multiple dimensions. The data includes time-stamped events, facilitating temporal analyses, and developing predictive models for patient outcomes. The features considered in this analysis were: heart rate, respiratory rate, saturation of peripheral oxygen (or SpO2), systolic arterial blood pressure, diastolic arterial blood pressure.

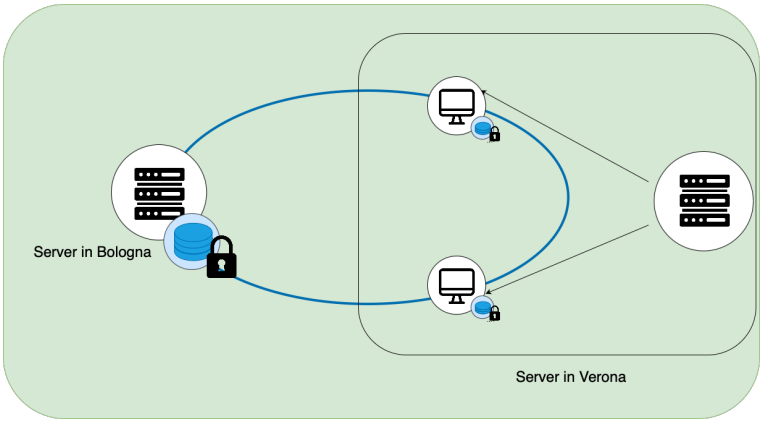


Figure 6.4: Schematic representation of swarm learning using the MIMIC III dataset.

6.3 Results

MNIST dataset

In Figure 6.5, we divided the original dataset into three equal proportions. We can see that the swarm model performs slightly better than the model of the single nodes. However, the central model, which is trained on the entire original dataset, outperforms the swarm model.

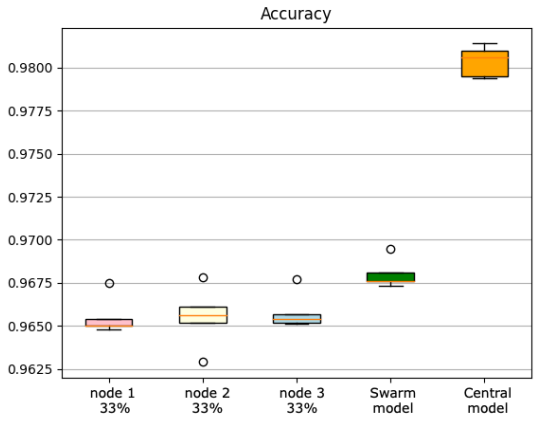


Figure 6.5: Boxplots of accuracy for each node, with portions 33% in the first node, 33% in the second node, and 33% in the third. Box plots are given by 10 different trainings using 10 random shuffles of the data with different random seeds. Orange line in within boxplot represents the mean; box limits, 1st and 3rd quartiles; remaining dots: outliers.

In Figure 6.6, we divided the original dataset in three non overlapping sets, in which we divided the digits in the different sets: in the first set, we put the digits $\{0,1,2\}$ in the first node, $\{3,4,5\}$ in the second, $\{6,7,8,9\}$ in the third. This figure shows the incredible power of swarm learning: even if the three different servers never see some of the digits during the training, the swarm model performs almost like it has seen all the digits. Thinking about real scenarios, even if a center or hospital never registers a specific patient sample in its dataset, thanks to a swarm model, the hospital's model would predict well new diagnoses of the same type.

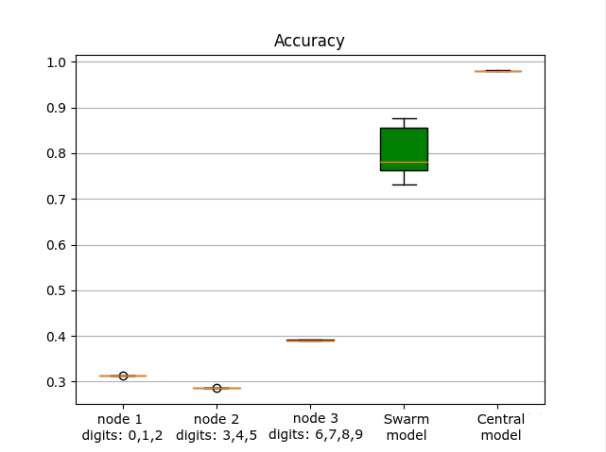
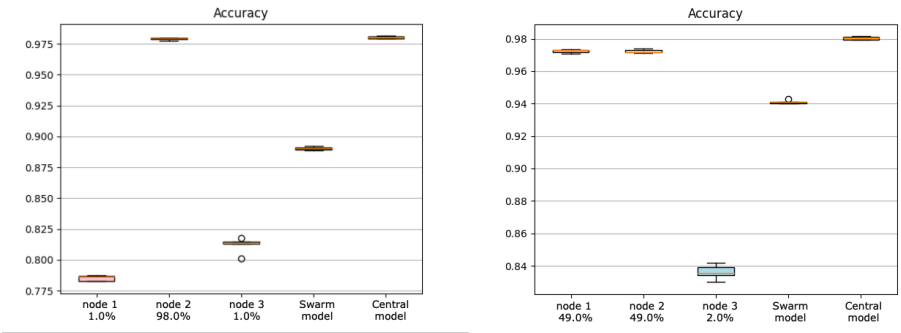


Figure 6.6: Boxplots of accuracy for each node, in which we divided the different classes of digits: {0,1,2} digits in the first node, {3,4,5} digits in the second node, and {6,7,8,9} in the third node. Box plots are given by 10 different trainings using 10 random shuffles of the data with different random seeds. The orange line within the boxplot represents the mean; box limits, 1st and 3rd quartiles; remaining dots: outliers.

In Figure 6.7a, we can see that we divided the dataset in different proportions. On the left, the proportions were: (1%,98%,1%). We can see that the local training with 98% of the training set outperforms the performance of the other two nodes, which is consistent with the fact that it had much more data. We can also see that the central model performs the same amount as for node 2, while the performance accuracy of the swarm node lies between the performances of the three nodes.

In Figure 6.7b, we can see a similar behavior, but in that case scenario, the proportions of the dataset were: (49%,49%,2%). Also, in this case, the swarm model performs worse than the first two nodes, which had the most significant proportions of the dataset. This indicates that the swarm model performs very well when the amount of data is uniform between the nodes, but when the training set is strongly unbalanced between the nodes, the swarm model performs worse than the other nodes. Notice also that the central model, which is trained on all the original datasets, is always the best.

Next, we ran the same experiments, but despite using simple average as a merging algorithm, we used *weighted average* this time. In this case, the swarm learning package allows setting a weight w with integer values with $w \in [0, 100]$ for the weighted mean. In Figure 6.8, we firstly set a value of $w = 90$ for the second node, which had proportion 98% of the dataset, and a value of $w = 1$ to the other nodes (Figure 6.8a). We can see



(a)

- First node: 1%
- Second node: 98%
- Third node: 1%

(b)

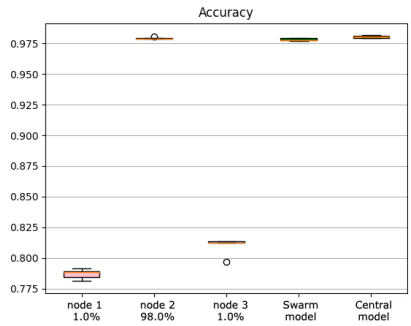
- First node: 49%
- Second node: 49%
- Third node: 2%

Figure 6.7: Box plots are given by 10 different trainings using 10 random data shuffles with different random seeds. The orange line within the boxplot represents the mean; box limits, 1st and 3rd quartiles; remaining dots: outliers.

that the swarm model performs almost like the second node, which is the node with the highest value of w. On the right instead (Figure 6.8b), we set a value of $w = 90$ for the third node, which had only 1% of the dataset. In this case, we can see that the swarm model follows the performance of the third node, leading to a rapid decrease in performance. We can see the same behaviour in Figure 6.9, where both in 6.9a and 6.9b, the performance of the swarm learning model follows the performance of the node with the highest weight.

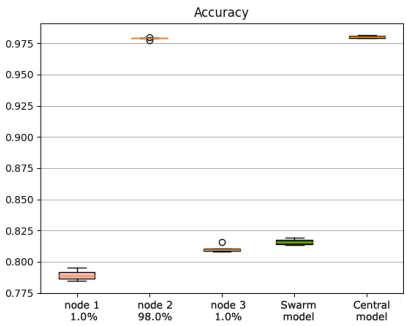
MIMIC dataset

In Figure 6.10, we can see that we divided the dataset in different proportions. On the left, the proportions were: (33%,33%,33%). We can see that the local training with 98% of the training set outperforms the performance of the other two nodes, which is consistent with the fact that it had much more data. We can also see that the central model performs the same amount as for node 2, while the performance accuracy of the swarm node lies between the performances of the three nodes. In Figure 6.10c, we can see similar behavior, but in that case scenario, the proportions of the dataset were: (49%,49%,2%). Also, in this case, the swarm model performs worse than the first two nodes, which had the most significant proportions of the dataset. This indicates that the swarm model performs very



(a)

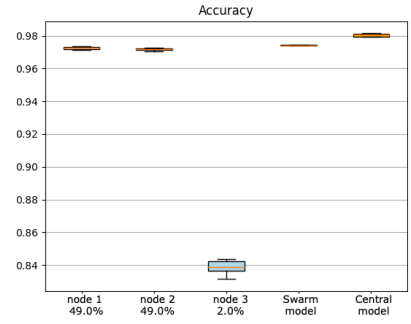
- Weight first node: 1
- Weight second node: 90
- Weight third node: 1



(b)

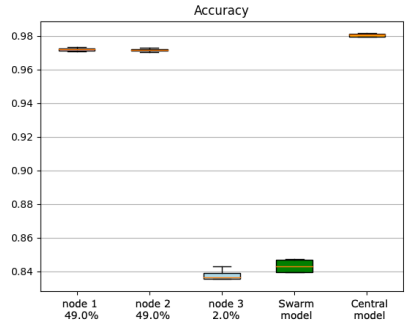
- Weight first node: 1
- Weight second node: 1
- Weight third node: 90

Figure 6.8: Box plots are given by 10 different trainings using 10 random data shuffles with different random seeds. The orange line within the boxplot represents the mean; box limits, 1st and 3rd quartiles; remaining dots: outliers.



(a)

- Weight first node: 1
- Weight second node: 90
- Weight third node: 1

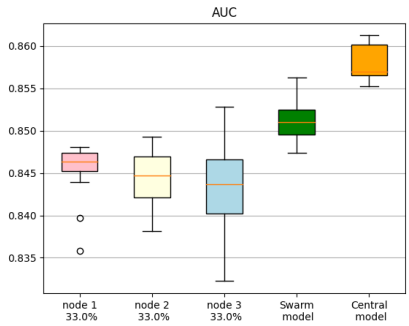


(b)

- Weight first node: 1
- Weight second node: 1
- Weight third node: 90

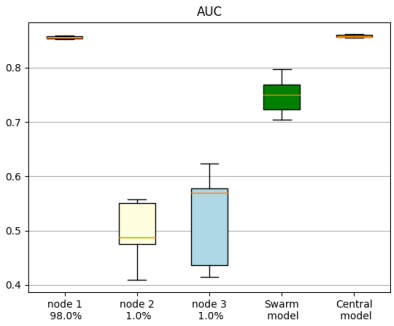
Figure 6.9: Box plots are given by 10 different trainings using 10 random data shuffles with different random seeds. The orange line within the boxplot represents the mean; box limits, 1st and 3rd quartiles.

well when the amount of data is uniform between the nodes, but when the training set is strongly unbalanced between the nodes, the swarm model performs worse than the other nodes. Notice also that the central model, which is the model trained on all the original datasets, is always the best.



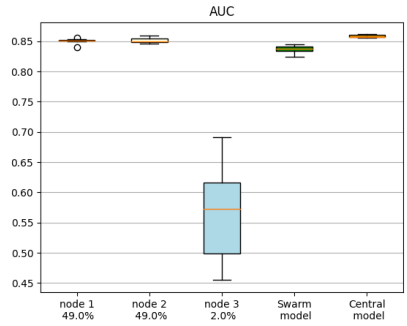
(a)

- First node: 33%
- Second node: 33%
- Third node: 33%



(b)

- First node: 98%
- Second node: 1%
- Third node: 1%



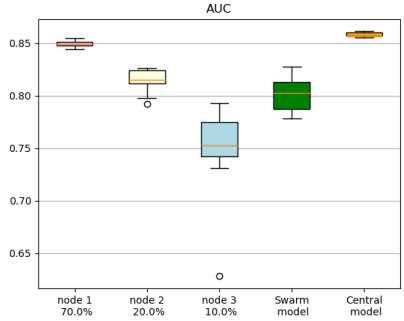
(c)

- First node: 49%
- Second node: 49%
- Third node: 2%

Figure 6.10: Box plots are given by 10 different trainings using 10 random shuffles of the data with different random seeds. Orange line in within boxplot represents the mean; box limits, 1st and 3rd quartiles; remaining dots: outliers.

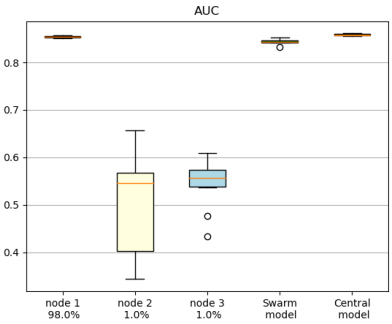
Next, also in these case, we repeated the same experiments but using weighted average as merging algorithm. This time, we set a value of w for each node that corresponds to the percentage of the training data of

that node; for example, if in one node there is 70% of the original dataset as training set, the value of the weight will be $w = 70$ (see Figure 6.11a). In Figure 6.11c and Figure 6.11b we can see that the performance of the swarm model follows the performance of the nodes which had the maximum percentage of the original dataset. This is coherent to the fact that the value of weight w reflects the amount of the data in that node.



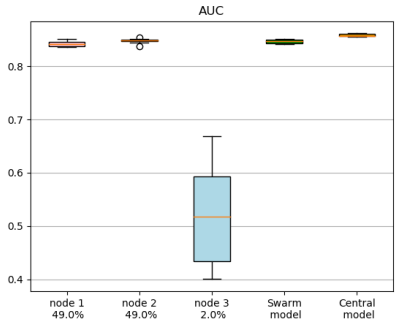
(a)

- First node: 70%
- Second node: 20%
- Third node: 10%



(b)

- First node: 98%
- Second node: 1%
- Third node: 1%



(c)

- First node: 49%
- Second node: 49%
- Third node: 2%

Figure 6.11: Box plots are given by 10 different trainings using 10 random shuffles of the data with different random seeds. The orange line within the boxplot represents the mean; box limits, 1st and 3rd quartiles; remaining dots: outliers.

6.4 Discussion

Acknowledging that swarm learning is a nascent tool in machine learning, it is essential to recognize that ongoing research and advancements are crucial for refining algorithms and optimizing the performance of the swarm model. In the published paper [Warnat-Herresthal et al., 2021] presenting swarm learning, authors investigated different configurations of the training, varying the number of samples between the nodes, and also different types of distributed data of heterogeneous diseases (COVID-19, tuberculosis, leukemia, and lung pathologies). However, they do not always compare SL-trained models with the central server model, where the model is trained on the whole dataset in a single server. Moreover, the authors used only a simple average as merging algorithm without analyzing different configurations with different weights for the nodes.

In addition to the published work on SL [Warnat-Herresthal et al., 2021], very few new works have been published on SL applications in the medical field [Saldanha et al., 2022, Saldanha et al., 2023]. In particular, both published works describe the application of SL on pathology images of gastric and colorectal cancer.

Similarly to the obtained results, Saldanha et al., in both of their works [Saldanha et al., 2022, Saldanha et al., 2023] showed that the central model metrics consistently outperform or, in some cases, perform on par with SL models. This indicates some limitations of this new technology.

However, despite this aspect, the capacity of SL to collaborate on model training without exchanging sensitive data presents a pragmatic solution for medicine. In a real world scenario involving several hospitals, sharing the parameters of a swarm model, the collective intelligence derived from diverse datasets becomes a driving force for improving artificial intelligence in medicine. As swarm learning matures, discoveries and algorithmic enhancements will play a pivotal role in maximizing the efficacy of this innovative approach in healthcare applications.

From a forward-looking perspective, further investigations focus on the assessment of the performance of the swarm learning technology through deeper analyses of different configurations of the framework, such as increasing the number of nodes, different proportions of the dataset within different nodes, and different weights of the nodes for the weighted average as merging algorithm. Moreover, further investigations will rely on implementing explainability techniques such as representative SHAP values and validation techniques within the swarm learning context.

Chapter 7

Conclusions

This work highlights the potential of artificial intelligence in healthcare, particularly in diagnosis and treatment. However, the successful integration of artificial intelligence into medical practices faces significant challenges that require immediate attention. The investigation has focused on three crucial aspects: explainability, reproducibility, and the scarcity of data due to privacy concerns.

Explainability is critical in building trust in AI systems, especially in medical applications where decisions directly impact patient well-being. For this reason, this work might give a contribution to the research domain proposing a new algorithm to compute *representative SHAP values*, average explanations which need to enhance reproducibility of explainability tools within different nested validation techniques. This approach contributes to the transparency and reliability of AI in medical decision-making, ensuring consistent explanations across various settings.

The scarcity of medical data, amplified by stringent privacy regulations, poses a significant obstacle to developing effective AI models. Swarm learning is an innovative solution which overcomes data scarcity issues and ensures compliance with privacy regulations, laying the foundation for developing more robust AI solutions in the medical domain. However, since swarm learning is a recently proposed technique, a sistematic analysis of its performance is lacking. To address this challenge, an exploration of its potential has began in this work, evaluating the performance of the swarm learning framework with different configurations of the dataset in a swarm network with 3 nodes. However, the performance of a central model trained with the whole dataset still outperform the swarm learning model. For this reason, further research may focus on the application of generative AI and data augmentation techniques to enhance the performance of the models.

This research underscores the importance of addressing key considerations such as explainability, reproducibility, and privacy concerns when deploying AI for healthcare applications. By tackling these challenges, the responsible and effective implementation of machine learning in the medical field is possible, ultimately enhancing patient care and advancing the potential of AI-driven solutions in healthcare.

Bibliography

[Adadi and Berrada, 2018] Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160.

[Agarap, 2018] Agarap, A. F. (2018). Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

[Al Zoubi et al., 2018] Al Zoubi, O., Ki Wong, C., Kuplicki, R. T., Yeh, H.-w., Mayeli, A., Refai, H., Paulus, M., and Bodurka, J. (2018). Predicting age from brain eeg signals—a machine learning approach. *Frontiers in Aging Neuroscience*, 10.

[Alonim et al., 2021] Alonim, H., Lieberman, I., Schayngesicht, G., and Tayar, D. (2021). *A retrospective study of prodromal variables associated with autism among a global group of infants during their first fifteen months of life*, 7. cited By 1.

[Altman, 1999] Altman, D. G. (1999). Statistics in the medical literature: 3. *Statistics in Medicine*, 18(4):487–490.

[Altomare et al., 2023] Altomare, D., Barkhof, F., Caprioglio, C., Collij, L., Scheltens, P., Lopes Alves, I., Bouwman, F., Berkhof, J., Van Maurik, I., Garibotto, V., Moro, C., Delrieu, J., Payoux, P., Saint-Aubert, L., Hitzel, A., Molinuevo, J., Grau-Rivera, O., Gispert, J., Drzezga, A., Jessen, F., Zeyen, P., Nordberg, A., Savitcheva, I., Jelic, V., Walker, Z., Edison, P., Demonet, J.-F., Gismondi, R., Farrar, G., Stephens, A., and Frisoni, G. (2023). Clinical effect of early vs late amyloid positron emission tomography in memory clinic patients: The amypad-dpms randomized clinical trial. *JAMA Neurology*, 80(6):548–557. cited By 3.

[Al'Aref et al., 2018] Al'Aref, S. J., Anchouche, K., Singh, G., Slomka, P. J., Kolli, K. K., Kumar, A., Pandey, M., Maliakal, G., van Rosendael, A. R., Beecy, A. N., Berman, D. S., Leipsic, J., Nieman, K., Andreini, D., Pontone, G., Schoepf, U. J., Shaw, L. J., Chang, H.-J., Narula, J., Bax, J. J.,

Guan, Y., and Min, J. K. (2018). Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. *European Heart Journal*, 40(24):1975–1986.

[Amir et al., 2021] Amir, S., van de Meent, J., and Wallace, B. C. (2021). On the impact of random seeds on the fairness of clinical classifiers. https://arxiv.org/abs/2104.06338.

[Antwarg et al., 2021] Antwarg, L., Miller, R. M., Shapira, B., and Rokach, L. (2021). Explaining anomalies detected by autoencoders using shapley additive explanations. *Expert Systems with Applications*, 186:115736.

[Aouedi et al., 2023] Aouedi, O., Sacco, A., Piamrat, K., and Marchetto, G. (2023). Handling privacy-sensitive medical data with federated learning: Challenges and future directions. *IEEE Journal of Biomedical and Health Informatics*, 27(2):790–803.

[Arrieta et al., 2020] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.

[Baron-Cohen, 1995] Baron-Cohen, S. (1995). *Mindblindness: An Essay on Autism and Theory of Mind*. cited By 4145.

[Baron-Cohen, 2000] Baron-Cohen, S. (2000). Theory of mind and autism: A review. *International Review of Research in Mental Retardation*, 23:169–184. cited By 419.

[Baron-Cohen et al., 1985] Baron-Cohen, S., Leslie, A., and Frith, U. (1985). Does the autistic child have a "theory of mind" ? *Cognition*, 21(1):37–46. cited By 5101.

[Bastos-Leite et al., 2007] Bastos-Leite, A., Van Der Flier, W., Van Straaten, E., Staekenborg, S., Scheltens, P., and Barkhof, F. (2007). The contribution of medial temporal lobe atrophy and vascular pathology to cognitive impairment in vascular dementia. *Stroke*, 38(12):3182–3185. cited By 97.

[Batunacun et al., 2021] Batunacun, Wieland, R., Lakes, T., and Nendel, C. (2021). Using shapley additive explanations to interpret extreme gradient boosting predictions of grassland degradation in xilingol, china. *Geoscientific Model Development*, 14(3):1493–1510.

[Beam et al., 2020] Beam, A. L., Manrai, A. K., and Ghassemi, M. (2020). Challenges to the Reproducibility of Machine Learning Models in Health Care. *JAMA*, 323(4):305–306.

[Beebe-Wang et al., 2021] Beebe-Wang, N., Okeson, A., Althoff, T., and Lee, S.-I. (2021). Efficient and explainable risk assessments for imminent dementia in an aging cohort study. *IEEE Journal of Biomedical and Health Informatics*, 25(7):2409–2420.

[Bi et al., 2020a] Bi, Y., Xiang, D., Ge, Z., Li, F., Jia, C., and Song, J. (2020a). An interpretable prediction model for identifying n7-methylguanosine sites based on xgboost and shap. *Molecular Therapy - Nucleic Acids*, 22:362–372.

[Bi et al., 2020b] Bi, Y., Xiang, D., Ge, Z., Li, F., Jia, C., and Song, J. (2020b). Prediction of the development of acute kidney injury following cardiac surgery by machine learning. *Crit Care*, 24.

[Bilello et al., 2015] Bilello, M., Doshi, J., Nabavizadeh, S., Toledo, J., Erus, G., Xie, S., Trojanowski, J., Han, X., and Davatzikos, C. (2015). Correlating cognitive decline with white matter lesion and brain atrophy magnetic resonance imaging measurements in alzheimer's disease. *Journal of Alzheimer's Disease*, 48(4):987–994. cited By 62.

[Blüthgen et al., 2021] Blüthgen, C., Patella, M., Euler, A., Baessler, B., Martini, K., von Spiczak, J., Schneiter, D., Opitz, I., and Frauenfelder, T. (2021). Computed tomography radiomics for the prediction of thymic epithelial tumor histology, tnm stage and myasthenia gravis. *PLOS ONE*, 16(12):1–16.

[Bracco et al., 2005] Bracco, L., Piccini, C., Moretti, M., Mascalchi, M., Sforza, A., Nacmias, B., Cellini, E., Bagnoli, S., and Sorbi, S. (2005). Alzheimer's disease: Role of size and location of white matter changes in determining cognitive deficits. *Dementia and Geriatric Cognitive Disorders*, 20(6):358–366. cited By 51.

[Burman, 1989] Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3):503–514.

[Caffarra et al., 2002] Caffarra, P., Vezzadini, G., Dieci, F., Zonato, F., and Venneri, A. (2002). A short version of the stroop test: Normative data in an italian population sample [una versione abbreviata del test di stroop:

Dati normativi nella popolazione italiana]. *Nuova Rivista di Neurologia*, 12(4):111–115. cited By 316.

[Chazette et al., 2021] Chazette, L., Brunotte, W., and Speith, T. (2021). Exploring explainability: a definition, a model, and a knowledge catalogue. In *2021 IEEE 29th international requirements engineering conference (RE)*, pages 197–208. IEEE.

[Chen et al., 2021] Chen, L., Song, J., Cheng, R., Wang, K., Liu, X., He, M., and Luo, T. (2021). Cortical thinning in the medial temporal lobe and precuneus is related to cognitive deficits in patients with subcortical ischemic vascular disease. *Frontiers in Aging Neuroscience*, 12. cited By 12.

[Chen and Guestrin, 2016a] Chen, T. and Guestrin, C. (2016a). Xgboost: A scalable tree boosting system. volume 13-17-August-2016, pages 785–794. cited By 20047.

[Chen and Guestrin, 2016b] Chen, T. and Guestrin, C. (2016b). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

[Chen et al., 2019] Chen, T., Xu, J., Ying, H., Chen, X., Feng, R., Fang, X., Gao, H., and Wu, J. (2019). Prediction of extubation failure for intensive care unit patients using light gradient boosting machine. *IEEE Access*, 7:150960–150968.

[Chishti et al., 2020] Chishti, S., Jaggi, K. R., Saini, A., Agarwal, G., and Ranjan, A. (2020). Artificial intelligence-based differential diagnosis: Development and validation of a probabilistic model to address lack of large-scale clinical datasets. *J Med Internet Res*, 22(4):e17550.

[Chlap et al., 2021] Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., and Haworth, A. (2021). A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 65(5):545–563.

[Ciobanu-Caraus et al., 2024] Ciobanu-Caraus, O., Aicher, A., Kernbach, J. M., Regli, L., Serra, C., and Staartjes, V. E. (2024). A critical moment in machine learning in medicine: on reproducible and interpretable learning. *Acta Neurochirurgica*, 166(1):14.

[Cirillo et al., 2020] 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. (2020). Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ digital medicine*, 3(1):1–11.

[Ciulli et al., 2016] Ciulli, S., Citi, L., Salvadori, E., Valenti, R., Poggesi, A., Inzitari, D., Mascalchi, M., Toschi, N., Pantoni, L., and Diciotti, S. (2016). Prediction of impaired performance in trail making test in mci patients with small vessel disease using dti data. *IEEE Journal of Biomedical and Health Informatics*, 20(4):1026–1033. cited By 27.

[Conti et al., 2015] Conti, S., Bonazzi, S., Laiacona, M., Masina, M., and Coralli, M. (2015). Montreal cognitive assessment (moca)-italian version: regression based norms and equivalent scores. *Neurological Sciences*, 36(2):209–214. cited By 165.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

[Daniels and Mandell, 2014] Daniels, A. and Mandell, D. (2014). Explaining differences in age at autism spectrum disorder diagnosis: A critical review. *Autism*, 18(5):583–597. cited By 412.

[Deb and Smith, 2021] Deb, D. and Smith, R. M. (2021). Application of random forest and shap tree explainer in exploring spatial (in)justice to aid urban planning. *ISPRS International Journal of Geo-Information*, 10(9).

[Deng, 2012] Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142.

[Di Martino and Delmastro, 2023] Di Martino, F. and Delmastro, F. (2023). Explainable ai for clinical and remote health applications: a survey on tabular and time series data. *Artificial Intelligence Review*, 56(6):5261–5315.

[Diciotti et al., 2013] Diciotti, S., Ciulli, S., Mascalchi, M., Giannelli, M., and Toschi, N. (2013). The "peeking" effect in supervised feature selection on diffusion tensor imaging data. *American Journal of Neuroradiology*, 34(9):E107. cited By 21.

[Dixon, 2006] Dixon, P. M. (2006). *Bootstrap Resampling*. John Wiley Sons, Ltd.

[Egle et al., 2022] Egle, M., Hilal, S., Tuladhar, A., Pirpamer, L., Hofer, E., Duering, M., Wason, J., Morris, R., Dichgans, M., Schmidt, R., Tozer, D., Chen, C., De Leeuw, F., and Markus, H. (2022). Prediction of dementia using diffusion tensor mri measures: the optimal collaboration. *Journal of Neurology, Neurosurgery and Psychiatry*, 93(1):14–23. cited By 14.

[El-Sappagh et al., 2021] El-Sappagh, S., Alonso, J. M., Islam, S., Sultan, A. M., and Kwak, K. S. (2021). A multilayer multimodal detection and prediction model based on explainable artificial intelligence for alzheimer's disease. *Scientific reports*, 11(1):1–26.

[Elder et al., 2017] Elder, J., Kreider, C., Brasher, S., and Ansell, M. (2017). Clinical impact of early diagnosis of autism on the prognosis and parent-child relationships. *Psychology Research and Behavior Management*, 10:283–292. cited By 138.

[Fan et al., 2021] Fan, Y., Shen, M., Huo, Y., Gao, X., Li, C., Zheng, R., and Zhang, J. (2021). Total cerebral small vessel disease burden on mri correlates with medial temporal lobe atrophy and cognitive performance in patients of a memory clinic. *Frontiers in Aging Neuroscience*, 13. cited By 5.

[Fawcett, 2006] Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874. cited By 14239.

[Fazekas et al., 1987] Fazekas, F., Chawluk, J., Alavi, A., Hurtig, H., and Zimmerman, R. (1987). Mr signal abnormalities at 1.5 t in alzheimer's dementia and normal aging. *American Journal of Roentgenology*, 149(2):351–356. cited By 3400.

[Feng et al., 2021] Feng, D.-C., Wang, W.-J., Mangalathu, S., and Taciroglu, E. (2021). Interpretable xgboost-shap machine-learning model for shear strength prediction of squat rc walls. *Journal of Structural Engineering*, 147(11):04021173.

[Fischl, 2012] Fischl, B. (2012). Freesurfer. *NeuroImage*, 62(2):774–781. 20 YEARS OF fMRI.

[Fischl et al., 2001] Fischl, B., Liu, A., and Dale, A. (2001). Automated manifold surgery: constructing geometrically accurate and topologically correct models of the human cerebral cortex. *IEEE Transactions on Medical Imaging*, 20(1):70–80.

[Franke and Gaser, 2019] Franke, K. and Gaser, C. (2019). Ten years of brainage as a neuroimaging biomarker of brain aging: What insights have we gained? *Frontiers in Neurology*, 10.

[Franz and Dawson, 2019] Franz, L. and Dawson, G. (2019). Implementing early intervention for autism spectrum disorder: A global perspective. *Pediatric Medicine*, 2. cited By 11.

[Free et al., 1996] Free, S., Sisodiya, S., Cook, M., Fish, D., and Shorvon, S. (1996). Three-dimensional fractal analysis of the white matter surface from magnetic resonance images of the human brain. *Cerebral Cortex*, 6(6):830–836.

[Gabbay-Dizdar et al., 2022] Gabbay-Dizdar, N., Ilan, M., Meiri, G., Faroy, M., Michaelovski, A., Flusser, H., Menashe, I., Koller, J., Zachor, D., and Dinstein, I. (2022). Early diagnosis of autism in the community is associated with marked improvement in social symptoms within 1–2 years. *Autism*, 26(6):1353–1363. cited By 28.

[Gallagher and Hutto, 2008] Gallagher, S. and Hutto, D. (2008). Understanding others through primary interaction and narrative practice. *The Shared Mind: Perspectives on Intersubjectivity*, pages 17–38. cited By 347.

[Gauthier et al., 2006] Gauthier, S., Reisberg, B., Zaudig, M., Petersen, R., Ritchie, K., Broich, K., Belleville, S., Brodaty, H., Bennett, D., Chertkow, H., Cummings, J., de Leon, M., Feldman, H., Ganguli, M., Hampel, H., Scheltens, P., Tierney, M., Whitehouse, P., and Winblad, B. (2006). Mild cognitive impairment. *Lancet*, 367(9518):1262–1270. cited By 2065.

[Ginestroni et al., 2009] Ginestroni, A., Battaglini, M., Della Nave, R., Moretti, M., Tessa, C., Giannelli, M., Caffarra, P., Nacmias, B., Bessi, V., Sorbi, S., Bracco, L., De Stefano, N., and Mascalchi, M. (2009). Early structural changes in individuals at risk of familial alzheimer's disease: A volumetry and magnetization transfer mr imaging study. *Journal of Neurology*, 256(6):925–932. cited By 30.

[Giorgio et al., 2019] Giorgio, A., Di Donato, I., De Leucio, A., Zhang, J., Salvadori, E., Poggesi, A., Diciotti, S., Cosottini, M., Ciulli, S., Inzitari, D., Pantoni, L., Mascalchi, M., Federico, A., Dotti, M., De Stefano, N., and on behalf of the VMCI-Tuscany Study Group (2019). Relevance of brain lesion location for cognition in vascular mild cognitive impairment. *NeuroImage: Clinical*, 22. cited By 10.

[Golomb et al., 1995] Golomb, J., Kluger, A., Gianutsos, J., Ferris, S., De Leon, M., and George, A. (1995). Nonspecific leukoencephalopathy associated with aging. *Neuroimaging Clinics of North America*, 5(1):33–44. cited By 43.

[Goñi et al., 2013] 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. (2013). Robust estimation of fractal measures for characterizing the structural complexity of the human brain: optimization and reproducibility. *Neuroimage*, 83:646–657.

[Gundersen and Kjensmo, 2018] Gundersen, O. E. and Kjensmo, S. (2018). State of the art: Reproducibility in artificial intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

[Haibe-Kains et al.,] Haibe-Kains, B., Adam, G. A., Hosny, A., and Khodakarami, F. Matters arising transparency and reproducibility in artificial intelligence. *John P. A. Ioannidis*, 10:34.

[Han et al., 2006] Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., Busa, E., Pacheco, J., Albert, M., Killiany, R., Maguire, P., Rosas, D., Makris, N., Dale, A., Dickerson, B., and Fischl, B. (2006). Reliability of mri-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. *NeuroImage*, 32(1):180–194.

[Hastie, 2013] Hastie, Tibshirani, F. (2013). *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. Springer.

[Heil et al., 2021] Heil, B. J., Hoffman, M. M., Markowetz, F., Lee, S.-I., Greene, C. S., and Hicks, S. C. (2021). Reproducibility standards for machine learning in the life sciences. *Nature Methods*, 18(10):1132–1135.

[Heinen et al., 2020] Heinen, R., Groeneveld, O., Barkhof, F., de Bresser, J., Exalto, L., Kuijf, H., Prins, N., Scheltens, P., van der Flier, W., Biessels, G., and behalf of the TRACE-VCI study group, O. (2020). Small vessel disease lesion type and brain atrophy: The role of co-occurring amyloid. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, 12(1). cited By 3.

[Holzinger, 2018] Holzinger, A. (2018). From machine learning to explainable ai. In *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, pages 55–66.

[Im et al., 2006] Im, K., Lee, J.-M., Yoon, U., Shin, Y.-W., Soon, B., In, Y., Juo, S., and Kim, S. (2006). Fractal dimension in human cortical surface: Multiple regression analysis with cortical thickness, sulcal depth, and folding area. *Human Brain Mapping*, 27(12):994–1003. cited By 148.

[Inzitari et al., 2009] Inzitari, D., Pracucci, G., Poggesi, A., Carlucci, G., Barkhof, F., Chabriat, H., Erkinjuntti, T., Fazekas, F., Ferro, J., Hennerici, M., Langhorne, P., O'Brien, J., Scheltens, P., Visser, M., Wahlund, L.-O., Waldemar, G., Wallin, A., and Pantoni, L. (2009). Changes in white matter as determinant of global functional decline in older independent outpatients: Three year follow-up of ladis (leukoaraiosis and disability) study cohort. *BMJ (Online)*, 339(7715):279–282. cited By 320.

[Jellinger, 2013] Jellinger, K. (2013). Pathology and pathogenesis of vascular cognitive impairment-a critical update. *Frontiers in Aging Neuroscience*, 5(APR). cited By 225.

[Johnson et al., 2016] Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

[Jokinen et al., 2015] Jokinen, H., Gonçalves, N., Vigário, R., Lipsanen, J., Fazekas, F., Schmidt, R., Barkhof, F., Madureira, S., Verdelho, A., Inzitari, D., Pantoni, L., Erkinjuntti, T., and Group, T. L. S. (2015). Early-stage white matter lesions detected by multispectral mri segmentation predict progressive cognitive decline. *Frontiers in Neuroscience*, 9(DEC). cited By 21.

[Jokinen et al., 2020] Jokinen, H., Koikkalainen, J., Laakso, H., Melkas, S., Nieminen, T., Brander, A., Korvenoja, A., Rueckert, D., Barkhof, F., Scheltens, P., Schmidt, R., Fazekas, F., Madureira, S., Verdelho, A., Wallin, A., Wahlund, L.-O., Waldemar, G., Chabriat, H., Hennerici, M., O'Brien, J., Inzitari, D., Lötjönen, J., Pantoni, L., and Erkinjuntti, T. (2020). Global burden of small vessel disease-related brain changes on mri predicts cognitive and functional decline. *Stroke*, 51(1):170–178. cited By 94.

[Jokinen et al., 2012] Jokinen, H., Lipsanen, J., Schmidt, R., Fazekas, F., Gouw, A., Van Der Flier, W., Barkhof, F., Madureira, S., Verdelho, A., Ferro, J., Wallin, A., Pantoni, L., Inzitari, D., and Erkinjuntti, T. (2012). Brain atrophy accelerates cognitive decline in cerebral small vessel disease the ladis study. *Neurology*, 78(22):1785–1792. cited By 120.

[Kang et al., 2012] Kang, X., Herron, T. J., Cate, A. D., Yund, E. W., and Woods, D. L. (2012). Hemispherically-unified surface maps of human cerebral cortex: Reliability and hemispheric asymmetries. *PLOS ONE*, 7(9):1–15.

[Keller et al., 2013] Keller, S. S., Ahrens, T., Mohammadi, S., Gerdes, J. S., Möddel, G., Kellinghaus, C., Kugel, H., Weber, B., Ringelstein, E. B., and Deppe, M. (2013). Voxel-based statistical analysis of fractional anisotropy and mean diffusivity in patients with unilateral temporal lobe epilepsy of unknown cause. *Journal of Neuroimaging*, 23(3):352–359.

[Kim, 2009] Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics Data Analysis*, 53(11):3735–3745.

[Kim and Kim, 2022] Kim, Y. and Kim, Y. (2022). Explainable heat-related mortality with random forest and shapley additive explanations (shap) models. *Sustainable Cities and Society*, 79:103677.

[King et al., 2009] King, R., George, A., Jeon, T., Hynan, L., Youn, T., Kennedy, D., and Dickerson, B. (2009). Characterization of atrophic changes in the cerebral cortex using fractal dimensional analysis. *Brain Imaging and Behavior*, 3(2):154–166. cited By 92.

[King, 2014] King, R. D. (2014). Computation of local fractal dimension values of the human cerebral cortex. *Applied Mathematics*, 2014.

[King et al., 2010] King, R. D., Brown, B., Hwang, M., Jeon, T., George, A. T., Initiative, A. D. N., et al. (2010). Fractal dimension analysis of the cortical ribbon in mild alzheimer's disease. *Neuroimage*, 53(2):471–479.

[Kingma and Ba, 2017] Kingma, D. P. and Ba, J. (2017). Adam: A method for stochastic optimization.

[Kiselev et al., 2003] Kiselev, V. G., Hahn, K. R., and Auer, D. P. (2003). Is the brain cortex a fractal? *Neuroimage*, 20(3):1765–1774.

[Kötter et al., 2001] Kötter, R., Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., Woods, R., Paus, T., Simpson, G., Pike, B., Holmes, C., Collins, L., Thompson, P., MacDonald, D., Iacoboni, M., Schormann, T., Amunts, K., Palomero-Gallagher, N., Geyer, S., Parsons, L., Narr, K., Kabani, N., Goualher, G. L., Boomsma, D., Cannon, T., Kawashima, R., and Mazoyer, B. (2001). A probabilistic atlas and reference system for the human brain: International consortium for brain

mapping (icbm). *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 356(1412):1293–1322.

[Lambert et al., 2016] Lambert, C., Benjamin, P., Zeestraten, E., Lawrence, A., Barrick, T., and Markus, H. (2016). Longitudinal patterns of leukoaraiosis and brain atrophy in symptomatic small vessel disease. *Brain*, 139(4):1136–1151. cited By 92.

[Lambert et al., 2015] Lambert, C., Sam Narean, J., Benjamin, P., Zeestraten, E., Barrick, T., and Markus, H. (2015). Characterising the grey matter correlates of leukoaraiosis in cerebral small vessel disease. *NeuroImage: Clinical*, 9:194–205. cited By 63.

[Landis and Koch, 1977] Landis, J. and Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174. cited By 54785.

[Lee et al., 2006] Lee, J. K., Lee, J.-M., Kim, J. S., Kim, I. Y., Evans, A. C., and Kim, S. I. (2006). A novel quantitative cross-validation of different cortical surface reconstruction algorithms using mri phantom. *NeuroImage*, 31(2):572–584.

[Lee Choong Ho, 2017] Lee Choong Ho, Y. H.-J. (2017). Medical big data: promise and challenges. *Kidney Res Clin Pract*, 36(1):3–11.

[Lemm et al., 2011] Lemm, S., Blankertz, B., Dickhaus, T., and Müller, K.-R. (2011). Introduction to machine learning for brain imaging. *NeuroImage*, 56(2):387–399. cited By 488.

[Li et al., 2020] Li, X., Yin, B., Tian, W., and Sun, Y. (2020). Performance of repeated cross validation for machine learning models in building energy analysis. In Wang, Z., Zhu, Y., Wang, F., Wang, P., Shen, C., and Liu, J., editors, *Proceedings of the 11th International Symposium on Heating, Ventilation and Air Conditioning (ISHVAC 2019)*, pages 523–531, Singapore. Springer Singapore.

[Loftus et al., 2022] Loftus, T. J., Tighe, P. J., Ozrazgat-Baslanti, T., Davis, J. P., Ruppert, M. M., Ren, Y., Shickel, B., Kamaleswaran, R., Hogan, W. R., Moorman, J. R., Upchurch, Jr, G. R., Rashidi, P., and Bihorac, A. (2022). Ideal algorithms in healthcare: Explainable, dynamic, precise, autonomous, fair, and reproducible. *PLOS Digital Health*, 1(1):1–16.

[Lombardi et al., 2022] Lombardi, A., Diacono, D., Amoroso, N., Biecek, P., Monaco, A., Bellantuono, L., Pantaleo, E., Logroscino, G., Blasi, R.,

Tangaro, S., et al. (2022). A robust framework to investigate the reliability and stability of explainable artificial intelligence markers of mild cognitive impairment and alzheimer's disease.

[Lombardi et al., 2021] Lombardi, A., Diacono, D., Amoroso, N., Monaco, A., Tavares, J. M. R. S., Bellotti, R., and Tangaro, S. (2021). Explainable deep learning for personalized age prediction with brain morphology. *Frontiers in Neuroscience*, 15.

[Lord, 1999] Lord, C. (1999). Autism diagnostic observation schedule. *(No Title)*.

[Lundberg and Lee, 2017a] Lundberg, S. and Lee, S.-I. (2017a). A unified approach to interpreting model predictions. *A Unified Approach to Interpreting Model Predictions*, pages 4765–4774. cited By 2238.

[Lundberg, 2018a] Lundberg, S. M. (2018a). Shap bar plot. https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/plots/bar.html.

[Lundberg, 2018b] Lundberg, S. M. (2018b). Shap beeswarm plot. https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/plots/beeswarm.html#A-simple-beeswarm-summary-plot.

[Lundberg, 2018c] Lundberg, S. M. (2018c). Shap explainer. https://shap.readthedocs.io/en/latest/example_notebooks/tabular_examples/tree_based_models/Python%20Version%20of%20Tree%20SHAP.html#Python-TreeExplainer.

[Lundberg et al., 2020] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):2522–5839.

[Lundberg and Lee, 2017b] Lundberg, S. M. and Lee, S.-I. (2017b). A unified approach to interpreting model predictions. https://arxiv.org/abs/1705.07874.

[Luyster et al., 2009] Luyster, R., Gotham, K., Guthrie, W., Coffing, M., Petrak, R., Pierce, K., Bishop, S., Esler, A., Hus, V., Oti, R., Richler, J., Risi, S., and Lord, C. (2009). The autism diagnostic observation schedule - toddler module: A new module of a standardized diagnostic measure for autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 39(9):1305–1320. cited By 314.

[Madan and Kensinger, 2018] Madan, C. and Kensinger, E. (2018). Predicting age from cortical structure across the lifespan. *European Journal of Neuroscience*, 47(5):399–416. cited By 55.

[Madan and Kensinger, 2016] Madan, C. R. and Kensinger, E. A. (2016). Cortical complexity as a measure of age-related brain atrophy. *NeuroImage*, 134:617–629.

[Marzi, 2023] Marzi, C. (2023). *chiaramarzi/fractalbrain-toolkit: fraxtalbrain-toolkit v1, 1. Zenodo*. cited By 1.

[Marzi et al., 2018] Marzi, C., Ciulli, S., Giannelli, M., Ginestroni, A., Tessa, C., Mascalchi, M., and Diciotti, S. (2018). Structural complexity of the cerebellum and cerebral cortex is reduced in spinocerebellar ataxia type 2. *Journal of Neuroimaging*, 28(6):688–693.

[Marzi et al., 2022] Marzi, C., Giannelli, M., Barucci, A., Tessa, C., and Mascalchi, M. (2022). Efficacy of mri data harmonization in the age of machine learning. a multicenter study across 36 datasets. *arXiv preprint arXiv:2211.04125*. cited By 1.

[Marzi et al., 2020] Marzi, C., Giannelli, M., Tessa, C., Mascalchi, M., and Diciotti, S. (2020). Toward a more reliable characterization of fractal properties of the cerebral cortex of healthy subjects during the lifespan. *Scientific Reports*, 10:16957.

[Marzi et al., 2021a] Marzi, C., Giannelli, M., Tessa, C., Mascalchi, M., and Diciotti, S. (2021a). Fractal analysis of mri data at 7 t: How much complex is the cerebral cortex? *IEEE Access*, 9:69226–69234.

[Marzi et al., 2021b] Marzi, C., Giannelli, M., Tessa, C., Mascalchi, M., and Diciotti, S. (2021b). Fractal analysis of mri data at 7 t: How much complex is the cerebral cortex? *IEEE Access*, 9:69226–69234. cited By 9.

[Mascalchi et al., 2013] Mascalchi, M., Ginestroni, A., Bessi, V., Toschi, N., Padiglioni, S., Ciulli, S., Tessa, C., Giannelli, M., Bracco, L., and Diciotti, S. (2013). Regional analysis of the magnetization transfer ratio of the brain in mild alzheimer disease and amnestic mild cognitive impairment. *American Journal of Neuroradiology*, 34(11):2098–2104. cited By 15.

[Mascalchi et al., 2014] Mascalchi, M., Ginestroni, A., Toschi, N., Poggesi, A., Cecchi, P., Salvadori, E., Tessa, C., Cosottini, M., De Stefano, N., Pracucci, G., Pantoni, L., Inzitari, D., Diciotti, S., Abbate, R., Bandinelli, M., Boddi, M., Cesari, F., Ciolli, L., Coppo, M., Bene, A., Giusti, B., Gori,

A., Nannucci, S., Pasi, M., Pescini, F., Valenti, R., Bonucelli, U., Chiti, A., Orlandi, G., Pagni, C., Siciliano, G., Tognoni, G., Federico, A., Stefano, N., Dotti, M., Formichi, P., Gambetti, C., Giorgio, A., Rossi, F., Stromillo, L., Zicari, E., Zolo, P., Tiezzi, A., Bertini, E., Brotini, S., Guidi, L., Lombardi, M., Mugnai, S., Notarelli, A., Bracco, L., Cadelo, M., Cisbani, R., Gabbani, L., Gori, G., Lambertucci, L., Massacesi, L., Mossello, E., Paganini, M., Piccininni, M., Pinto, F., Pozzi, C., Sorbi, S., Zaccara, G., Borgogni, T., Mancuso, M., Marconi, R., Mazzoni, M., Vista, M., Meucci, G., Bellini, G., Gabrielli, L., Frittelli, C., Galli, R., Gambaccini, G., Bartolini, S., Biagini, C., Caleri, V., Vanni, P., Calvani, D., Giorgi, C., Magnolfi, S., Palumbo, P., Valente, C., Rossi, A., Tassi, R., Boschi, S., and Baldacci, F. (2014). The burden of microstructural damage modulates cortical activation in elderly subjects with mci and leuko-araiosis. a dti and fmri study. *Human Brain Mapping*, 35(3):819–830. cited By 46.

[McMahan et al., 2017] McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

[Miller, 2019] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

[Moassefi et al., 2023] Moassefi, M., Rouzrokh, P., Conte, G. M., Vahdati, S., Fu, T., Tahmasebi, A., Younis, M., Farahani, K., Gentili, A., Kline, T., et al. (2023). Reproducibility of deep learning algorithms developed for medical imaging analysis: A systematic review. *Journal of Digital Imaging*, 36(5):2306–2312.

[Molinaro et al., 2005] Molinaro, A., Simon, R., and Pfeiffer, R. (2005). Prediction error estimation: A comparison of resampling methods. *Bioinformatics*, 21(15):3301–3307. cited By 937.

[Monti et al., 2020] Monti, R. P., Gibberd, A., Roy, S., Nunes, M., Lorenz, R., Leech, R., Ogawa, T., Kawanabe, M., and Hyvärinen, A. (2020). Interpretable brain age prediction using linear latent variable models of functional connectivity. *PLOS ONE*, 15(6):1–25.

[Mueller and Guido, 2017] Mueller, A. and Guido, S. (2017). *Introduction to machine Learning with Python: A guide for Data Scientists*. O'Reilly Media.

[Murtaza et al., 2023] Murtaza, H., Ahmed, M., Khan, N. F., Murtaza, G., Zafar, S., and Bano, A. (2023). Synthetic data generation: State of the art in health care domain. *Computer Science Review*, 48:100546.

[Nasreddine et al., 2005] Nasreddine, Z., Phillips, N., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J., and Chertkow, H. (2005). The montreal cognitive assessment, moca: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699. cited By 14691.

[Nazlee et al., 2023] Nazlee, N., Waiter, G., and Sandu, A.-L. (2023). Age-associated sex and asymmetry differentiation in hemispheric and lobar cortical ribbon complexity across adulthood: A uk biobank imaging study. *Human Brain Mapping*, 44(1):49–65. cited By 4.

[Nocentini et al., 2006] Nocentini, U., Giordano, A., Di Vincenzo, S., Panella, M., and Pasqualetti, P. (2006). The symbol digit modalities test - oral version: Italian normative data. *Functional Neurology*, 21(2):93–96. cited By 77.

[Nooner et al., 2012] 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. (2012). The nki-rockland sample: a model for accelerating the pace of discovery science in psychiatry. *Frontiers in neuroscience*, 6:152.

[O'Sullivan, 2008] O'Sullivan, M. (2008). Leukoaraiosis. *Practical Neurology*, 8(1):26–38. cited By 126.

[Pani et al., 2022a] Pani, J., Marzi, C., Stensvold, D., Wisløff, U., Håberg, A. K., and Diciotti, S. (2022a). Longitudinal study of the effect of a 5-year exercise intervention on structural brain complexity in older adults. a generation 100 substudy. *NeuroImage*, page 119226.

[Pani et al., 2022b] Pani, J., Marzi, C., Stensvold, D., Wisløff, U., Håberg, A. K., and Diciotti, S. (2022b). Longitudinal study of the effect of a 5-year exercise intervention on structural brain complexity in older adults. a generation 100 substudy. *NeuroImage*, 256:119226.

[Pantoni et al., 2019a] Pantoni, L., Marzi, C., Poggesi, A., Giorgio, A., De Stefano, N., Mascalchi, M., Inzitari, D., Salvadori, E., and Diciotti, S. (2019a). 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: Clinical*, 24:101990.

[Pantoni et al., 2019b] Pantoni, L., Marzi, C., Poggesi, A., Giorgio, A., De Stefano, N., Mascalchi, M., Inzitari, D., Salvadori, E., and Diciotti, S. (2019b). 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: Clinical*, 24. cited By 27.

[Paolucci, 2020] Paolucci, C. (2020). A radical enactivist approach to social cognition. *Perspectives in Pragmatics, Philosophy and Psychology*, 23:59–74. cited By 6.

[Paolucci, 2021] Paolucci, C. (2021). *Cognitive semiotics: Integrating signs, minds, meaning and cognition*. cited By 35.

[Parsa et al., 2020] Parsa, A. B., Movahedi, A., Taghipour, H., Derrible, S., and Mohammadian, A. K. (2020). Toward safer highways, application of xgboost and shap for real-time accident detection and feature analysis. *Accident Analysis Prevention*, 136:105405.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830. cited By 50414.

[Pineau et al., 2021] Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Larochelle, H. (2021). Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *J. Mach. Learn. Res.*, 22(1).

[Poggesi et al., 2012] Poggesi, A., Salvadori, E., Pantoni, L., Pracucci, G., Cesari, F., Chiti, A., Ciolli, L., Cosottini, M., Del Bene, A., De Stefano, N., Diciotti, S., Dotti, M., Ginestroni, A., Giusti, B., Gori, A., Nannucci, S., Orlandi, G., Pescini, F., Valenti, R., Abbate, R., Federico, A., Mascalchi, M., Murri, L., and Inzitari, D. (2012). Risk and determinants of dementia in patients with mild cognitive impairment and brain subcortical vascular changes: A study of clinical, neuroimaging, and biological markersthe vmci-tuscany study: Rationale, design, and methodology. *International Journal of Alzheimer's Disease*. cited By 29.

[Rajpurkar et al.,] Rajpurkar, P., Chen, E., Oishi, B., Banerjee, O., and Topol, E. J. Ai in health and medicine.

[Ribeiro et al., 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938.

[Rieke et al., 2020] Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., et al. (2020). The future of digital health with federated learning. *NPJ digital medicine*, 3(1):119.

[Rodríguez-Pérez and Bajorath, 2020] Rodríguez-Pérez, R. and Bajorath, J. (2020). Interpretation of compound activity predictions from complex machine learning models using local approximations and shapley values. *Journal of Medicinal Chemistry*, 63(16):8761–8777. PMID: 31512867.

[Rosas et al., 2002] Rosas, H., Liu, A., Hersch, S., Glessner, M., Ferrante, R., Salat, D., van Der Kouwe, A., Jenkins, B., Dale, A., and Fischl, B. (2002). Regional and progressive thinning of the cortical ribbon in huntington's disease. *Neurology*, 58(5):695–701.

[Rotholz et al., 2017] Rotholz, D., Kinsman, A., Lacy, K., and Charles, J. (2017). Improving early identification and intervention for children at risk for autism spectrum disorder. *Pediatrics*, 139(2). cited By 69.

[Sabuncu, 2020] Sabuncu, M. R. (2020). Intelligence plays dice: Stochasticity is essential for machine learning. https://arxiv.org/abs/2008.07496.

[Sala et al., 1992] Sala, S., Laiacona, M., Spinnler, H., and Ubezio, C. (1992). A cancellation test: Its reliability in assessing attentional deficits in alzheimer's disease. *Psychological Medicine*, 22(4):885–901. cited By 161.

[Salari et al., 2022] Salari, N., Rasoulpoor, S., Rasoulpoor, S., Shohaimi, S., Jafarpour, S., Abdoli, N., Khaledi-Paveh, B., and Mohammadi, M. (2022). The global prevalence of autism spectrum disorder: a comprehensive systematic review and meta-analysis. *Italian Journal of Pediatrics*, 48(1). cited By 50.

[Saldanha et al., 2023] Saldanha, O. L., Muti, H. S., Grabsch, H. I., Langer, R., Dislich, B., Kohlruss, M., Keller, G., van Treeck, M., Hewitt, K. J., Kolbinger, F. R., et al. (2023). Direct prediction of genetic aberrations from pathology images in gastric cancer with swarm learning. *Gastric cancer*, 26(2):264–274.

[Saldanha et al., 2022] Saldanha, O. L., Quirke, P., West, N. P., James, J. A., Loughrey, M. B., Grabsch, H. I., Salto-Tellez, M., Alwers, E., Cifci, D., Ghaffari Laleh, N., et al. (2022). Swarm learning for decentralized artificial intelligence in cancer histopathology. *Nature Medicine*, 28(6):1232–1239.

[Salvadori et al., 2016] Salvadori, E., Poggesi, A., Valenti, R., Pracucci, G., Pescini, F., Pasi, M., Nannucci, S., Marini, S., Del Bene, A., Ciolli, L., Ginestroni, A., Diciotti, S., Orlandi, G., Di Donato, I., De Stefano, N., Cosottini, M., Chiti, A., Federico, A., Dotti, M., Bonuccelli, U., Inzitari, D., Pantoni, L., Abbate, R., Boddi, M., Cesari, F., Coppo, M., Giusti, B., Gori, A., Mascalchi, M., Cecchi, P., Pagni, C., Siciliano, G., Tognoni, G., Formichi, P., Gambetti, C., Giorgio, A., Rossi, F., Stromillo, L., Zicari, E., Zolo, P., Tiezzi, A., Bertini, E., Brotini, S., Guidi, L., Lombardi, M., Mugnai, S., Notarelli, A., Bracco, L., Cadelo, M., Cisbani, R., Gabbani, L., Gori, G., Lambertucci, L., Massacesi, L., Mossello, E., Paganini, M., Piccininni, M., Pinto, F., Pozzi, C., Sorbi, S., Zaccara, G., Borgogni, T., Mancuso, M., Marconi, R., Mazzoni, M., Vista, M., Meucci, G., Bellini, G., Gabrielli, L., Frittelli, C., Galli, R., Gambaccini, G., Bartolini, S., Biagini, C., Caleri, V., Vanni, P., Calvani, D., Giorgi, C., Magnolfi, S., Palumbo, P., Valente, C., Rossi, A., Tassi, R., Boschi, S., Baldacci, F., and Group, V.-T. S. (2016). Operationalizing mild cognitive impairment criteria in small vessel disease: The vmci-tuscany study. *Alzheimer's and Dementia*, 12(4):407–418. cited By 31.

[Sandu et al., 2014a] Sandu, A.-L., Izard, E., Specht, K., Beneventi, H., Lundervold, A., and Ystad, M. (2014a). Post-adolescent developmental changes in cortical complexity. *Behavioral and Brain Functions*, 10(1). cited By 23.

[Sandu et al., 2008a] Sandu, A.-L., Rasmussen Jr., I.-A., Lundervold, A., Kreuder, F., Neckelmann, G., Hugdahl, K., and Specht, K. (2008a). Fractal dimension analysis of mr images reveals grey matter structure irregularities in schizophrenia. *Computerized Medical Imaging and Graphics*, 32(2):150–158. cited By 65.

[Sandu et al., 2008b] Sandu, A.-L., Specht, K., Beneventi, H., Lundervold, A., and Hugdahl, K. (2008b). Sex-differences in grey-white matter structure in normal-reading and dyslexic adolescents. *Neuroscience Letters*, 438(1):80–84. cited By 34.

[Sandu et al., 2014b] Sandu, A.-L., Staff, R., McNeil, C., Mustafa, N., Ahearn, T., Whalley, L., and Murray, A. (2014b). Structural brain com-

plexity and cognitive decline in late life - a longitudinal study in the aberdeen 1936 birth cohort. *NeuroImage*, 100:558–563. cited By 35.

[Sandu et al., 2022] Sandu, A.-L., Waiter, G., Staff, R., Nazlee, N., Habota, T., McNeil, C., Chapko, D., Williams, J., Fall, C., Chandak, G., Pene, S., Krishna, M., McIntosh, A., Whalley, H., Kumaran, K., Krishnaveni, G., and Murray, A. (2022). Sexual dimorphism in the relationship between brain complexity, volume and general intelligence (g): a cross-cohort study. *Scientific Reports*, 12(1). cited By 3.

[Saw and Ng, 2022] Saw, S. N. and Ng, K. H. (2022). Current challenges of implementing artificial intelligence in medical imaging. *Physica Medica*, 100:12–17.

[Schaer et al., 2008] Schaer, M., Cuadra, M. B., Tamarit, L., Lazeyras, F., Eliez, S., and Thiran, J.-P. (2008). A surface-based approach to quantify local cortical gyrification. *IEEE Transactions on Medical Imaging*, 27(2):161–170.

[Scheda and Diciotti, 2022] Scheda, R. and Diciotti, S. (2022). Explanations of machine learning models in repeated nested cross-validation: An application in age prediction using brain complexity features. *Applied Sciences (Switzerland)*, 12(13). cited By 9.

[Selvaraju et al., 2017] 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*, pages 618–626.

[Shaikhina and Khovanova, 2017] Shaikhina, T. and Khovanova, N. A. (2017). Handling limited datasets with neural networks in medical applications: A small-data approach. *Artificial Intelligence in Medicine*, 75:51–63.

[Shapley, 1952] Shapley, L. S. (1952). *17. A Value for n-Person Games*, pages 307–318. Princeton University Press.

[Shehab et al., 2022] Shehab, M., Abualigah, L., Shambour, Q., Abu-Hashem, M. A., Shambour, M. K. Y., Alsalibi, A. I., and Gandomi, A. H. (2022). Machine learning in medical applications: A review of state-of-the-art methods. *Computers in Biology and Medicine*, 145:105458.

[Shrikumar et al., 2017] Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. *CoRR*, abs/1704.02685.

[Siciarz et al., 2021] Siciarz, P., Alfaifi, S., Uytven, E. V., Rathod, S., Koul, R., and McCurdy, B. (2021). Machine learning for dose-volume histogram based clinical decision-making support system in radiation therapy plans for brain tumors. *Clinical and Translational Radiation Oncology*, 31:50–57.

[Simonyan et al., 2013] Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

[Stower, 2020] Stower, H. (2020). Transparency in medical ai. *Nature Medicine*, 26:14–16.

[Sun et al., 2022] Sun, W., Huang, L., Cheng, Y., Qin, R., Xu, H., Shao, P., Ma, J., Yao, Z., Shi, L., and Xu, Y. (2022). Medial temporal atrophy contributes to cognitive impairment in cerebral small vessel disease. *Frontiers in Neurology*, 13. cited By 2.

[Teitelbaum et al., 1998] Teitelbaum, P., Teitelbaum, O., Nye, J., Fryman, J., and Maurer, R. (1998). Movement analysis in infancy may be useful for early diagnosis of autism. *Proceedings of the National Academy of Sciences of the United States of America*, 95(23):13982–13987. cited By 516.

[Trevarthen and Hubley, 1978] Trevarthen, C. and Hubley, P. (1978). Secondary intersubjectivity: Confidence, confiding and acts of meaning in the first year. *Action, Gesture and Symbol: The Emergence of Language*, pages 183–229. cited By 743.

[Valenti et al., 2016] Valenti, R., Del Bene, A., Poggesi, A., Ginestroni, A., Salvadori, E., Pracucci, G., Ciolli, L., Marini, S., Nannucci, S., Pasi, M., Pescini, F., Diciotti, S., Orlandi, G., Cosottini, M., Chiti, A., Mascalchi, M., Bonuccelli, U., Inzitari, D., and Pantoni, L. (2016). Cerebral microbleeds in patients with mild cognitive impairment and small vessel disease: The vascular mild cognitive impairment (vmci)-tuscany study. *Journal of the Neurological Sciences*, 368:195–202. cited By 26.

[van 't Hof et al., 2021] van 't Hof, M., Tisseur, C., van Berckelear-Onnes, I., van Nieuwenhuyzen, A., Daniels, A., Deen, M., Hoek, H., and Ester,

W. (2021). Age at autism spectrum disorder diagnosis: A systematic review and meta-analysis from 2012 to 2019. *Autism*, 25(4):862–873. cited By 142.

[Vanwinckelen and Blockeel, 2012] Vanwinckelen, G. and Blockeel, H. (2012). On estimating model accuracy with repeated cross-validation. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjc5dre_pf1AhUtNOwKHUpQClcQFnoECBEQAQ&url=https%3A%2F%2Flirias.kuleuven.be%2Fretrieve%2F186558%2F&usg=AOvVaw3sAhjDtQOB2NwGcalWuwpk.

[Volkmar, 2014] Volkmar, F. (2014). Editorial: The importance of early intervention. *Journal of Autism and Developmental Disorders*, 44(12):2979–2980. cited By 22.

[Walsh et al., 2021] Walsh, I., Fishman, D., Garcia-Gasulla, D., Titma, T., Pollastri, G., Capriotti, E., Casadio, R., Capella-Gutierrez, S., Cirillo, D., Conte, A. D., Dimopoulos, A. C., Angel, V. D. D., Dopazo, J., Fariselli, P., Fernández, J. M., Huber, F., Kreshuk, A., Lenaerts, T., Martelli, P. L., Navarro, A., Broin, P. , Pinero, J., Piovesan, D., Reczko, M., Ronzano, F., Satagopam, V., Savojardo, C., Spiwok, V., Tangaro, M. A., Tartari, G., Salgado, D., Valencia, A., Zambelli, F., Harrow, J., Psomopoulos, F. E., and Tosatto, S. C. E. (2021). Dome: recommendations for supervised machine learning validation in biology. *Nature Methods*.

[Wang et al., 2021] Wang, K., Tian, J., Zheng, C., Yang, H., Ren, J., Liu, Y., Han, Q., and Zhang, Y. (2021). Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and shap. *Computers in Biology and Medicine*, 137:104813.

[Warnat-Herresthal et al., 2021] Warnat-Herresthal, S., Schultze, H., Shastry, K. L., Manamohan, S., Mukherjee, S., Garg, V., Sarveswara, R., Händler, K., Pickkers, P., Aziz, N. A., et al. (2021). Swarm learning for decentralized and confidential clinical machine learning. *Nature*, 594(7862):265–270.

[Williams et al., 2019] Williams, O., Zeestraten, E., Benjamin, P., Lambert, C., Lawrence, A., Mackinnon, A., Morris, R., Markus, H., Barrick, T., and Charlton, R. (2019). Predicting dementia in cerebral small vessel disease using an automatic diffusion tensor image segmentation technique. *Stroke*, 50(10):2775–2782. cited By 12.

[Williams et al., 2017] Williams, O., Zeestraten, E., Benjamin, P., Lambert, C., Lawrence, A., Mackinnon, A., Morris, R., Markus, H., Charlton, R., and Barrick, T. (2017). Diffusion tensor image segmentation of the cerebrum provides a single measure of cerebral small vessel disease severity related to cognitive change. *NeuroImage: Clinical*, 16:330–342. cited By 26.

[Wu et al., 2019] Wu, A., Sharrett, A. R., Gottesman, R. F., Power, M. C., Mosley, T. H., Jack, C. R., Knopman, D. S., Windham, B. G., Gross, A. L., and Coresh, J. (2019). Association of brain magnetic resonance imaging signs with cognitive outcomes in persons with nonimpaired cognition and mild cognitive impairment. *JAMA network open*, 2(5):e193359–e193359.

[Yagis et al., 2021] Yagis, E., Atnafu, S. W., García Seco de Herrera, A., Marzi, C., Scheda, R., Giannelli, M., Tessa, C., Citi, L., and Diciotti, S. (2021). Effect of data leakage in brain mri classification using 2d convolutional neural networks. *Scientific reports*, 11(1):1–13.

[Ye et al., 2015] Ye, B. S., Seo, S. W., Kim, J.-H., Kim, G. H., Cho, H., Noh, Y., Kim, H. J., Yoon, C. W., Woo, S.-y., Kim, S. H., et al. (2015). Effects of amyloid and vascular markers on cognitive decline in subcortical vascular dementia. *Neurology*, 85(19):1687–1693.

[Youden, 1950] Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1):32–35.

[Zeestraten et al., 2017] Zeestraten, E. A., Lawrence, A. J., Lambert, C., Benjamin, P., Brookes, R. L., Mackinnon, A. D., Morris, R. G., Barrick, T. R., and Markus, H. S. (2017). Change in multimodal mri markers predicts dementia risk in cerebral small vessel disease. *Neurology*, 89(18):1869–1876.

[Zhang et al., 2021] Zhang, X., Yao, L., Wang, X., Monaghan, J., McAlpine, D., and Zhang, Y. (2021). A survey on deep learning-based non-invasive brain signals: recent advances and new frontiers. *Journal of Neural Engineering*, 18(3):031002.

[Zhao et al., 2022] Zhao, K., Duka, B., Xie, H., Oathes, D. J., Calhoun, V., and Zhang, Y. (2022). A dynamic graph convolutional neural network framework reveals new insights into connectome dysfunctions in adhd. *NeuroImage*, 246:118774.