From local counterfactuals to global feature importance: efficient, robust, and model-agnostic explanations for brain connectivity networks

Antonio Luca Alfeo, Antonio G. Zippo, Vincenzo Catrambone, Mario G.C.A. Cimino, Nicola Toschi, Gaetano Valenza

This is a preprint. Please cite using:

```
@article{from2023,
    author={Alfeo, Antonio Luca and Zippo, Antonio G. and Catrambone, Vincenzo and Cimino, Mario G.C.A. and Toschi, Nicola and Vale
    title={From local counterfactuals to global feature importance: efficient, robust, and model-agnostic explanations for brain connectivity n
    journal={Computer Methods and Programs in Biomedicine},
    year={2023},
    volume={236},
    pages={107550},
    publisher={Elsevier BV},
    doi={10.1016/j.cmpb.2023.107550},
    issn={0169-2607},
}
```

Antonio Luca Alfeo, Antonio G. Zippo, Vincenzo Catrambone, Mario G.C.A. Cimino, Nicola Toschi, Gaetano Valenza. "From local counterfactuals to global feature importance: efficient, robust, and model-agnostic explanations for brain connectivity networks" Computer Methods and Programs in Biomedicine 236 (2023): 107550.

From local counterfactuals to global feature importance: efficient, robust, and model-agnostic explanations for brain connectivity networks

Antonio Luca Alfeo[a,b,*], Antonio G. Zippo[c], Vincenzo Catrambone[a,b], Mario G.C.A. Cimino[a,b], Nicola Toschi[d] and Gaetano Valenza[a,b]

[a]Department of Information Engineering, University of Pisa, Largo Lucio Lazzarino, 1, Pisa, 56126, Italy
[b]Bioengineering & Robotics Research Center E. Piaggio, University of Pisa, Largo Lucio Lazzarino, 1, Pisa, 56126, Italy
[c]Institute of Neuroscience, Consiglio Nazionale delle Ricerche, Via Raoul Follereau, 3, Vedano al Lambro (MB), 20854, Italy
[d]Department of Biomedicine and Prevention, University of Rome Tor Vergata, Via Montpellier 1, Roma, 00133, Italy

ARTICLE INFO

Keywords:
eXplainable Artificial Intelligence
fMRI
Affective Computing
Feature Importance
Counterfactual explanation

ABSTRACT

Background: Explainable artificial intelligence (XAI) is a technology that can enhance trust in mental state classifications by providing explanations for the reasoning behind artificial intelligence (AI) models outputs, especially for high-dimensional and highly-correlated brain signals. Feature importance and counterfactual explanations are two common approaches to generate these explanations, but both have drawbacks. While feature importance methods, such as shapley additive explanations (SHAP), can be computationally expensive and sensitive to feature correlation, counterfactual explanations only explain a single outcome instead of the entire model.

Methods: To overcome these limitations, we propose a new procedure for computing global feature importance that involves aggregating local counterfactual explanations. This approach is specifically tailored to fMRI signals and is based on the hypothesis that instances close to the decision boundary and their counterfactuals mainly differ in the features identified as most important for the downstream classification task. We refer to this proposed feature importance measure as Boundary Crossing Solo Ratio (BoCSoR), since it quantifies the frequency with which a change in each feature in isolation leads to a change in classification outcome, i.e., the crossing of the model's decision boundary.

Results and Conclusions: Experimental results on synthetic data and real publicly available fMRI data from the Human Connect project show that the proposed BoCSoR measure is more robust to feature correlation and less computationally expensive than state-of-the-art methods. Additionally, it is equally effective in providing an explanation for the behavior of any AI model for brain signals. These properties are crucial for medical decision support systems, where many different features are often extracted from the same physiological measures and a gold standard is absent. Consequently, computing feature importance may become computationally expensive, and there may be a high probability of mutual correlation among features, leading to unreliable results from state-of-the-art XAI methods.

1. Introduction

Recent advances in machine learning (ML), and in particular in artificial intelligence (AI), have shown great potential for a variety of applications in the biomedical field, including protein folding, protein design, molecular medicine [1], as well as in the analysis and classification of physiological data, including brain signals [2; 3]. However, AI models are often criticized for their *black box* nature, which means that their inner workings are not transparent and are difficult to interpret. The increasing complexity of these models has led to a need for explainable artificial intelligence (XAI) algorithms that can provide insight into the reasoning behind the output of these models [4; 5; 6]. In the context of physiological data analysis and classification, especially considering brain signals, XAI can be particularly valuable. Brain signals are highly complex and often highly correlated, which makes it challenging to extract meaningful features and understand the underlying physiological processes that contribute to specific patterns. XAI can provide valuable insights into the mechanisms underlying these patterns and help researchers and clinicians better understand and interpret the results of physiological data analysis and classification [1].

XAI has recently gained significant attention as a potential tool for advancing neuroscience research [7; 8; 9]. XAI approaches have been employed to compare multi-modal brain data, behavioral and computational data, as well as stimulus descriptions [10]. More recently, XAI techniques have been successfully applied to longitudinally monitor subjects affected by mild cognitive impairment [11], to study factors contributing to stroke prediction [12], to highlight key features in epilepsy detection systems [13], and to shed light on brain dynamics associated with the aging process [14]. These studies showcase the versatility of XAI techniques in investigating brain signals and highlight the potential of XAI in facilitating the development of effective and accurate diagnostic and therapeutic tools.

*Corresponding author
luca.alfeo@unipi.it (A.L. Alfeo)
ORCID(s): 0000-0002-0928-3188 (A.L. Alfeo)

1.1. Explainable Artificial Intelligence

In recent years, there has been a growing interest in the development of XAI algorithms, which aim to provide clear and interpretable explanations of the reasoning behind AI models' predictions. XAI can help improve the transparency and trustworthiness of AI models and make them more accessible to non-experts. To illustrate, XAI technology can help medical personnel better justify medical treatment for patients, as well as take advantage of what the AI system has learned from the data to gain new scientific insights [15].

XAI approaches attach the so-called explanations for the reasoning of the AI algorithm to the provided predictions, allowing domain experts (e.g., biomedical scientists) to validate and trust the algorithms in a much smoother manner [16]. This view is supported by recent regulations on personal data processing (i.e., the General Data Protection Regulation [17]), which include some level of explainability as a requirement to be met in order to employ AI in real-world clinical decision-making processes [18].

The explanations generated by XAI approaches can be characterized by different properties. First, the explanations can be *local* or *global*. Local explanations motivate the classification outcome of a given instance, while global explanations provide insight into the whole model. Moreover, the explanations can come in different forms and shapes[19]: (i) *rule-based* explanations approximate the decision process embedded in the algorithm by associating labels to the thresholds of the input features [20]; (ii) *instance-based* explanations associate a labeled instance to some prototypes or counterexamples to trigger similarity-based reasoning in the end-user (e.g., the domain expert) [21]; (iii) *input-based* explanations explain the behavior of an AI model by grading the contribution of each input feature to the prediction [22]. Choosing the best explanation strategy is a design choice that depends on how comprehensibility and faithfulness are valued within a given application. Typically, comprehensible (i.e., compact, unambiguous) explanations are not faithful (i.e., comprehensively describing the AI model), and vice versa [23].

In a medical decision process, the end-users are not AI experts, thus the explanation needs to be as comprehensible as possible. To this aim, the explanation form that can be employed are the input-based (e.g., feature importance) and instance-based (e.g., counterfactual) [23].

Feature importance is one of the most widely used explanation forms, possibly due to the availability of model-agnostic approaches that can generate a feature ranking [22]. For example, the Shapley additive explanations (SHAP, [24]) framework is considered a gold standard among XAI approaches due to its solid theoretical background and wide applicability [25]. SHAP feature importance [26], estimates how important a feature is by measuring the average marginal contribution of a feature across all the possible combinations of features. This measure is computed for every data instance and then aggregated to provide a single ranking. Recently, the extensive use of SHAP has exposed its limitations, such as computational complexity, which grows exponentially with the number of features [27], and its sensitivity to correlation among features [28]. Unfortunately, both these conditions often occur in biomedical data [29]. Finally, SHAP does not provide insight into the behavior of the model with unseen instances [27]. Exemplarily, authors in [30] employed SHAP feature importance approaches to drive an algorithm for generating counterfactuals by modifying the value of the most important features.

To address this issue, counterfactual explanations can be employed. Intuitively, given a data instance i and its predicted class C_i, a counterfactual is an instance c 'similar' to i that has been allocated to a different predicted class ($C_c != C_i$). A counterfactual explanation is based on finding that 'similar' instance, meaning examining the minimum change that will result in a change in the predicted class. Unfortunately, the definition of 'minimum change' is not univocal. In some cases, it is considered the minimum number of features to change, in others, the minimum distance between the original instance and the counterfactual instances [31]. A counterfactual explanation can be found by *(i)* using a "brute force procedure", i.e. specifying the step size and the ranges of values for each feature to be explored around the instance being explained [32]; *(ii)* employing a specific loss measure and solving an optimization problem [33]; or *(iii)* adopting a heuristic search strategy, e.g., searching within a reference population of instances to be used as counterfactuals [34]. According to the results in [31], the latter strategies have the smallest computational cost (i.e., one or two orders of magnitude) as compared to the previous ones, and are often based on K-Nearest Neighbour procedures [34]. The main limitation of counterfactual explanations is the fact that they are instance-specific, i.e. no general information about the model reasoning as a whole is extracted [31; 35]. The authors in [36] propose a causality-based XAI approach based on probabilistic contrasting counterfactuals to generate global, local, and contextual explanations. However, the model requires structured knowledge, such as a causal graph, and does not provide an actual feature importance measure. The authors in [37] use a variational autoencoder approach to generate local explanations for an AI approach by approximating the decision boundary in the neighborhood of an instance to be explained. Anchor [38] produces local decision rules that are consistent with the decision boundary. Similarly, local surrogates [39] focuse on the decision boundary by generating instances in a hypersphere around the point to be explained, which is not feasible for large datasets.

1.2. In this work

In this paper, we propose a new XAI method, especially suitable for the analysis of brain signals. We employ the term *XAI method* to describe any process that can offer insights into how an ML model processes data instances to yield classifications. In our proposed approach, the generated insights pertain to the importance of each feature for the ML model. The most widely used method for determining the feature importance is SHAP. However, the literature reveals that

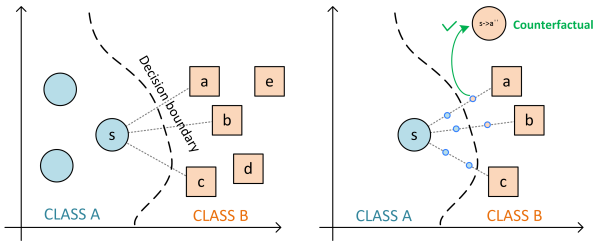

Figure 1: Example of the procedure for finding the minimally-different counterfactual for instance *s*, belonging to class A. The three nearest neighbors of *s* from the counterfactual class (i.e., class B) are considered. For each of them, two equally distant midpoints are generated. Among all the midpoints corresponding to a classification outcome equal to B, the closest is considered as the minimally-different counterfactual.

SHAP can be computationally demanding, and its reliability may be affected by feature correlation [28]. Our innovative approach to measuring feature importance for ML models that process tabular data relies on local counterfactuals, addressing these limitations. To assess the effectiveness of our proposed approach, we conducted experiments using the publicly available fMRI dataset Human Connectome Project (HCP) [40]. Our results demonstrate that our approach provides interpretable explanations for the model's decisions. In contrast to previous counterfactual XAI approaches, such as those proposed by Vlassopoulos et al. [37], Laugel et al. [39], and Ribeiro et al. [38], our proposed approach provides global explanations, addressing a major limitation of these methods. Furthermore, our results on both synthetic and publicly available real-world datasets demonstrate that our proposed feature importance measure is more robust to feature correlation and less computationally expensive than SHAP, while remaining capable of providing a thorough explanation of the behavior of any AI model for tabular data.

The paper is structured as follows. Section 2 presents the proposed XAI approach. Section 3 details the experimental dataset, while section 4 explains the experimental setup and the obtained results. Finally, section 5 and 6 discuss the results and outline the conclusions, respectively.

2. The Proposed Boundary Crossing Solo Ratio (BoCSoR) XAI algorithm

To the best of our knowledge, the proposed approach (i.e., BoCSoR) is among the few methods that integrate feature importance and counterfactual explanations. BoCSoR is based on two key assumptions: *(i)* feature importance indicates the most critical features for identifying a class, i.e., distinguishing it from other classes, and *(ii)* the counterfactuals of an instance are the most similar instances assigned to a different class, which lie beyond the decision boundary. Our hypothesis is that by considering instances near the decision boundary [37], the boundary is more likely to be crossed when the most important features are modified [41]. More formally, given the original class O and the counterfactual class C, we define B as the set of boundary instances b, as the instances of class O with a distance to their nearest neighbour of class C (i.e. c_{nn_b}) smaller than a certain percentile (th) of the distances obtained with all the instances o of class O (Eq. 1).

$$\begin{aligned} D &= \{dist(o, c_{nn_o}) \; \forall o \in O\} \\ B &= \{b \in O, \; dist(b, c_{nn_b}) < percentile(th, D)\} \end{aligned} \tag{1}$$

Given a boundary instance $b \in B$, the corresponding minimally-different counterfactual $closestCF_b$ is a instance recognized as class C characterized by minimal distance from b (Eq. 2). More on the concept of minimality employed in our approach in Algorithm 1.

$$closestCF_b \in C, \; dist(b, closestCF_b) \; is \; minimal \tag{2}$$

The feature f_i^b at index i is *relevant* if by substituting its value in $closestCF_b$ with its value in b, $closestCF_b$ is classified as class O. The importance of the feature at index i is the occurrence with which f_i is *relevant* considering all the boundary instances (Eq. 3).

$$BoCSoR_{f_i} = |\{f_i^b \; is \; relevant \; \forall b \in B\}| \tag{3}$$

In the following, we present the implementation of the proposed approach via pseudo-code. For all the procedures detailed in this Section, the Euclidean distance is considered as the reference distance measure (Eq. 4). In Eq. 4, a and b are two exemplary instances consisting of N features, whereas f_i^a and f_i^b are the values of the i^{th} feature for a and b, respectively.

$$dist(a, b) = \sqrt{\sum_{i=1}^{N} (f_i^a - f_i^b)^2} \tag{4}$$

According to [31], efficient approaches for counterfactual search can be based on K-Nearest Neighbor (NN) procedures. In essence, for the instance to be explained, the closest instances belonging to a different class can be utilized as potential counterfactuals. In Algorithm 1, the NN search is conducted using the method *KNNfromClass* (line 2) in which s denotes the instance, and k represents the number of closest instances of class c to be found. However, the nearest instance of another class may not correspond to the minimal change needed to achieve a different classification. To address this issue, midpoints are generated between each potential counterfactual and the original instance. In Algorithm 1 this is accomplished via the method *intermediatePointsBetween* (line 4) which provides a number of evenly spaced instances between s and c equal to $steps$. We collect each midpoint identified as an instance of the counterfactual class (lines 6-8 of Algorithm 1). This process is repeated for every potential counterfactual (see Fig. 1). We then calculate the distance between the collected midpoints and the original

instance, and the closest one is selected as the counterfactual that minimally differs from s (lines 11-13 of Algorithm 1). In our approach, we employ the concept of 'minimality' in a relative sense, referring to the step-wise exploration of the space between two instances belonging to different classes. This is rather than in an absolute sense, which would indicate the minimum distance necessary to alter the classification outcome. While our approach sacrifices the guarantee of achieving an absolute minimum distance, it substantially reduces computational costs, making it a more practical and favourable choice.

Algorithm 1 Pseudocode of the proposed procedure to efficiently obtain the closest counterfactual for a given instance (i.e., $findCF$).

Requires:
$M \Leftarrow$ trained machine learning model
$s \Leftarrow$ instance of which a counterfactual needs to be found
$class_s \Leftarrow$ class of s
$class_c \Leftarrow$ counterfactual class
$k \Leftarrow$ # closest neighbours of s from $class_c$
$steps \Leftarrow$ # intermediate steps between s and its neighbours
Procedure:

1: $explanations \Leftarrow emptyList()$
2: $counterfactuals \Leftarrow KNNfromClass(s, k, class_c)$
3: **for each** $c \in counterfactuals$ **do**
4: $\quad points \Leftarrow intermediatePointsBetween(s, c, steps)$
5: $\quad$ **for each** $p \in points$ **do**
6: $\quad\quad$ **if** $M.predict(p) == class_c$ **then**
7: $\quad\quad\quad explanations.append(p)$
8: $\quad\quad$ **end if**
9: $\quad$ **end for**
10: **end for**
11: $explDist \Leftarrow computeDistance(s, explanations)$
12: $minDist \Leftarrow min(explDist)$
13: $closestCF \Leftarrow select(explanations, explDist == minDist)$
14: **return** $closestCF$

After identifying the minimally different counterfactual using Algorithm 1, the subsequent step involves determining the features that can potentially cross the decision boundary when modified in isolation. To accomplish this, the value of each feature of the counterfactual instance is replaced with the corresponding value from the original instance, thereby generating a new instance. Initially, the index and value of each feature for the closest counterfactual are collected (line 3 of Algorithm 2). Then, a new instance is created by replacing the value of the feature at index i for the closest counterfactual instance with the value of the same feature in the original instance (line 5 of Algorithm 2). If this new instance yields a different classification outcome compared to the counterfactual instance, that feature is deemed relevant (lines 6-8 of Algorithm 2). Finally, the procedure outlined in Algorithm 2 gathers all relevant features for a given instance.

Our proposed feature importance measure, dubbed Boundary Crossing Solo Ratio (BoCSoR), is determined by the

Algorithm 2 Pseudocode of the procedure to obtain the relevant feature for a single instance to cross the decision boundary (i.e., $relevantFeatures$).

Requires:
$M \Leftarrow$ trained machine learning model
$s \Leftarrow$ instance of which a counterfactual needs to be found
$class_s \Leftarrow$ class of s
$class_c \Leftarrow$ counterfactual class
$k \Leftarrow$ # closest neighbours of s from $class_c$
$steps \Leftarrow$ # intermediate steps between s and its neighbours
Procedure:

1: $relevantFeaures \Leftarrow emptyList()$
2: $closestCF \Leftarrow findCF(M, s, k, steps, class_s, class_c)$
3: $[index, value] \Leftarrow valueChangeByFeature(s, closestCF)$
4: **for each** $i \in index$ **do**
5: $\quad CF_i \Leftarrow changeFeature(index_i, value_i)$
6: $\quad$ **if** $M.predict(CF_i) == class_s$ **then**
7: $\quad\quad relevantFeatures.append(index_i)$
8: $\quad$ **end if**
9: **end for**
10: **return** $relevantFeatures$

frequency at which each feature is considered relevant by the procedure outlined in Algorithm 2 when applied to instances close to the decision boundary. Closeness here, is defined as the inter-class distance, i.e. the Euclidean distance (Eq. 4) between the instance and its nearest instance from another class. We select instances with an inter-class distance smaller than a given percentile among all inter-class distances for the boundary instances. To find these, we first compute the pairwise (Euclidean) distance between all features of the original and counterfactual classes (lines 2-4 of Algorithm 3), and then select only the instances with a distance smaller than the given percentile (lines 5-6 of Algorithm 3). Finally, the relevant features (Algorithm 2) for each boundary instance are aggregated (lines 7-11 of Algorithm 3). BoCSoR measures the frequency at which a single change of each feature (i.e., with other features unchanged) results in crossing the decision boundary. To determine this measure, we replace the value of a feature at a given index for each boundary instance with the value of the same feature in the original instance (line 5 of Algorithm 2). We then compare the resulting instance with the corresponding counterfactual instance and consider the feature relevant if the classification outcome is different (lines 6-8 of Algorithm 2). In summary, BoCSoR considers the frequency at which each feature can result in crossing the decision boundary for instances close to it (Algorithm 3).

We present the time complexity analysis of the aforementioned procedures considering the number of instances N, the number of features f, and the main parameters of the procedures, i.e. k, $steps$, and $percentileTh$. To reduce the required number of distance computations and ensure a quick NN search, a variety of tree-based data structures have been proposed. Among those, we employ the so-called *ball tree*. According to the official *scikit-learn* documentation,

Algorithm 3 Pseudocode of the procedure to obtain the BoCSoR measure for a given decision boundary (i.e., $BoCSoR$).

Requires:
$M \Leftarrow$ trained machine learning model
$S \Leftarrow$ set of all the data instances
$class_o \Leftarrow$ original class
$class_c \Leftarrow$ counterfactual class
$percentileTh \Leftarrow$ threshold of the data
$k \Leftarrow$ # closest neighbours of s from $class_c$
$steps \Leftarrow$ # intermediate steps between s and its neighbours
Procedure:

$switches \Leftarrow emptyList()$
$set_O \Leftarrow select(data, label == class_o)$
$set_C \Leftarrow select(data, label == class_c)$
$pairwiseDist \Leftarrow computeDistance(set_O, set_C)$
$th \Leftarrow percentile(pairwiseDist, percentileTh)$
$instancesToExplain \Leftarrow select(set_C, pairwiseDist < th)$
for each $s \in instancesToExplain$ **do**
    $RF \Leftarrow relevantFeatures(M, s, k, steps, class_o, class_c)$
    $switches.append(RF)$
end for
$featureImportance \Leftarrow frequenceByFeature(switches)$
return $featureImportance$

the query time with a ball tree grows as approximately $O(f \cdot log(N))$, where f is the number of features and N is the number of instances. In Algorithm 1, the *NN* search query is computed k times, generating k potential counterfactuals ($O(k \cdot f \cdot log(N))$. For each one of the k potential counterfactuals, a number of *steps* midpoints are generated. Each one of those is used to produce a classification via the ML model. It results in a total amount of $k \cdot steps$ operations with a constant complexity ($O(k \cdot steps)$). It follows the computation of the distances between the instance s and the midpoints ($O(k \cdot steps)$) and the search for the minimum (worst case, $O(k \cdot steps)$). Overall, Algorithm 1 results in a complexity equal to $O(k \cdot f \cdot log(N) + 3 \cdot k \cdot steps)$, that can be simplified to $O(k \cdot f \cdot log(N) + k \cdot steps)$.

Algorithm 2 exploits Algorithm 1 to find the minimally-different counterfactual. Then, it switches the values of each feature and employs the obtained instance to compute a classification via the ML model. These constant complexity operations are repeated for each feature ($O(f)$). Overall, Algorithm 2 results in a complexity equal to $O(k \cdot f \cdot log(N) + k \cdot steps + f)$.

Algorithm 3 exploits Algorithm 2 for each boundary instance of class s, that is $percentileTh$ percentage of all the instances of the class. Assuming a balanced classification problem, in which each class has N instances, both the pairwise distance computation and the selection operation result in $O(N^2)$ complexity each. Those are followed by the executions of Algorithm 2. Thus, Algorithm 3 results in a final complexity equal to $O(2 \cdot N^2 + N \cdot \text{percentileTh}(k \cdot f \cdot log(N \cdot percentileTh) + k \cdot steps + f))$.

Since the value of $percentileTh$ is bounded, we can simplify the time complexity of Algorithm 3 as $O(N^2 + N \cdot (k \cdot f \cdot log(N) + k \cdot steps + f))$. If we also consider as bounded the parameterizations k and $steps$ (e.g. equal to 10), we can again simplify the time complexity, firstly as $O(N^2 + N(f \cdot log(N) + f))$ and then as per Eq. 5.

$$timeComplexity_{BoCSoR} = O(N^2 + N \cdot f \cdot log(N)) \quad (5)$$

Compared to SHAP, which has a time complexity that increases linearly with the number of instances and exponentially with the number of features [24], our proposed approach results in significantly smaller time complexity.

3. Experimental Data and Setup

3.1. Synthetic Datasets

In order to evaluate the reliability of feature importance measures, we constructed a tabular dataset (Dataset1) with known feature importance, following the methodology described in [42]. We used the *make_classification* method from the Python library *scikit-learn* [43] to generate normally distributed clusters of points around the vertices of a hypercube defined by five features. These features were interdependent, and the dataset was contaminated with noise to create a set of informative features followed by redundant features, i.e., linear combinations of informative ones. The amount of Gaussian noise added to each redundant feature increased linearly, resulting in a data set with high feature importance for informative features and predicted linearly decreasing importance along redundant features. The data generation procedure allowed us to vary the number of features, the number of informative features, and the number of instances. Further details on the data generation procedure can be found in [44].

By applying different feature importance approaches to this dataset, we could examine the extent to which the computed importance measures aligned with the imposed importance. We could also evaluate the impact of feature correlation on the reliability of feature importance computation.

To further explore the impact of feature correlation on feature importance measures, we created another synthetic dataset (Dataset2). This dataset was generated using the same procedure as Dataset1 but with the two most important and the two least important features duplicated to replace the other four features, thereby introducing high correlation between the least and most important features. This allowed us to assess the effect of the correlation between features with different importance levels on the computation of feature importance measures.

The ground truth feature importance for Dataset1 and Dataset2 does not consist of numerical values, which renders it unsuitable for measuring feature importance assessment error. Nonetheless, the method used to generate the synthetic

dataset provides a relative measure of feature importance, comparable to other studies that generate data with known feature importance [45]. Specifically, the synthetic dataset reveals which features are replicated (and thus most correlated) or informative (and thus most important) and which features become less important due to the progressive addition of noise. However, it is impossible to measure how the addition of noise may decrease feature importance in absolute terms. Despite this limitation, if two XAI approaches provide different importance measures for the same feature, it is feasible to investigate which measure is closest to the known informativeness level of the features, examine the conditions behind this difference, and verify if it consistently occurs with features of varying levels of informativeness. For instance, we anticipated that the average importance measure for the top informative features should surpass the average importance measure for the least informative features. Therefore, in our experimentation, we focused on groups of features characterized by opposite levels of importance and/or maximum correlation with other features in the dataset. Our quantitative results were aimed at exploring the relationship between the known relative feature importance and the importance measures provided by SHAP and BoCSoR.

3.2. Real Datasets: fMRI data from HCP

The Human Connectome Project (HCP) is a consortium of the National Institutes of Health of the United States that recruited participants for large-scale studies on human brain's anatomical and functional connectivity. The HCP consortium collected both resting state and task-evoked activities mostly from healthy subjects. So far, the largest study published is the Young Adults $HCP-1200$ which includes about 1200 healthy participants (aged $22-35$ y) who performed several tasks over 2 fMRI trials. In the present study, to test the usability and performance of the proposed method, the analysis was performed on Social and Emotional processing task data. For all subjects, all fMRI volumes were registered to a common reference space ($Brainnettome, http://www.brainnetome.org/$) parcelled into 123 cortical and subcortical regions per each hemisphere (246 in total, for ROI details, see [46]). More specifically, the considered recognition tasks are:

- Emotion Processing task [47]: participants watch either a fearful human face (12 different, 6 per gender, 2 trials) or a simple meaningless shape on a display for 18 seconds (3 seconds per trial) with no inter-stimulus intervals. The exact timing of stimuli presentation randomly changes from subject to subject with a standard deviation of about 0.13 seconds across 1200 subjects in order to avoid habituation.

- Social task [48]: 12 silent animated shapes (i.e., big red and small blue triangles) were shown on a screen during MRI acquisition and they are designed to mimic social interactions. Two kinds of animations are considered: ToM (3 for each trial) with animations eliciting mental state attributions, and RA (2 for each trial) comprehending animations of randomly moving shapes. Each animation lasts 23 seconds.

All subject-wise regions of interest (ROI) time-series were averaged within each ROI and preprocessed into subject-wise, ROI-to-ROI adjacency matrices, which has been proven to be successful in previous studies [49; 50; 51; 52; 53; 54], calculated as Pearson's correlation coefficient for each possible pair of the 246 ROIs. This approach generates a database with 30135 columns and more than a thousand rows ($((246 * 245)/2)$., which constituted our final dataset to be used for classification.

The groups considered are listed as follows: Superior Frontal Gyrus (SFG), Middle Frontal Gyrus (MFG), Inferior Frontal Gyrus (IFG), Orbital Gyrus (OrG), Precentral Gyrus (PrG), Paracentral Lobule (PCL), Superior Temporal Gyrus (STG), Middle Temporal Gyrus (MTG), Inferior Temporal Gyrus (ITG), Fusiform Gyrus (FuG), Parahippocampal Gyrus (PhG), posterior Superior Temporal Sulcus (pSTS), Superior Parietal Lobule (SPL), Inferior Parietal Lobule (IPL), Precuneus (Pcun), Postcentral Gyrus (PoG), Insular Gyrus (INS), Cingulate Gyrus (CG), MedioVentral Occipital Cortex (MVOcC), Lateral Occipital Cortex (LOcC), Amygdala (Amyg), Hippocampus (Hipp), Basal Ganglia (BG), Thalamus (Tha).

4. Experimental Results

The proposed approach, BoCSoR, aims to provide a measure of feature importance. While SHAP feature importance estimates are widely regarded as the gold standard in XAI literature, we investigated whether BoCSoR can offer improved computational efficiency and increased robustness to feature correlations - two of the main issues with SHAP [28]. To do this, we employed SHAP as a baseline method and evaluated the level of agreement between SHAP and BoCSoR. Additionally, we tested whether any disagreement between the two methods was more likely to occur with features that had high correlations with other features. As a model-agnostic approach, BoCSoR can be applied to any approach that processes tabular data. Our focus in this study was to explain the classification model rather than to achieve the best classification accuracy. Therefore, we employed shallow classifiers provided by *scikit-learn* [43] for all our experiments. Specifically, we used:

- Catboost [55], a ML approach that employs a gradient boosting approach to build an ensemble of decision trees;

- Multi-layer Perceptron (MLP) [56], a ML approach based on fully connected neural networks;

- Gaussian Process Classifier (GPC) [57], a kernel-based ML approach (like Support Vector Machine), aimed at predicting highly calibrated class membership probabilities;

Table 1
Wall-clock time [s] to compute the feature importance measure. Average ± standard deviation.

Data	SHAP	BoCSoR
Dataset1	41.72 ± 4.95	**1.28 ± 0.09**
Dataset2	39.08 ± 2.31	**1.23 ± 0.15**
HCP - Emotion	2604.2 ± 148.36	**185.1 ± 24.654**
HCP - Social	2618.3 ± 150.89	**169.58 ± 13.16**

It is important to note that our study was not focused on achieving the best possible classification performance. Therefore, we used the default hyperparameters provided by *scikit-learn* [43] for each of these approaches.

To ensure reliable results, all experimental evaluations were conducted via a 10-fold Monte Carlo cross-validation scheme.

We also compared the wall-clock time of BoCSoR and SHAP. To generate a fair comparison, we used the same number of instances considered by BoCSoR, i.e., the instances close to the decision boundary. To this end, we employed a $KernelExplainer$ in SHAP and employed different subsampling strategies. We conducted all computations on a hardware platform with a CPU Intel Core $i7-6700$ at $2.60-3.50GHz$, $6M$ cache, and $16GB$ DDR3L $1600MHz$ RAM. The wall-clock time for each method is reported in Table 1. This comparison provides evidence of the superiority of BoCSoR in terms of the wall-clock time (lower is better).

4.1. Synthetic Datasets

We computed the feature importance for Dataset1 using both SHAP and BoCSoR with the three ML approaches listed at the beginning of Section 4 (Fig. 2). The solid line represents the average feature importance measure provided by BoCSoR, while the dotted line represents the average feature importance measure provided by SHAP. The background color signifies the ideal trend of feature importance according to the ground truth, with the first five features considered important (green background) and the subsequent features becoming less important due to the addition of noise (transitioning from light green to red). The resulting average feature importance measures display a consistent overall descending trend across different measures (BoCSoR and SHAP) and ML approaches.

To assess the agreement between SHAP and BoCSoR, we computed Pearson's correlation coefficient between their feature importance values, as illustrated in Figure 3. To ensure a fair comparison, we rescaled the importance values between 0 and 1 using a min-max normalization. All correlation coefficients obtained were significant ($p < 0.05$), indicating a positive and significant correlation between BoCSoR and SHAP. This result was consistent across all three ML approaches used in our study.

However, while there is a high correlation between BoCSoR and SHAP importance values, discrepancies can occur in some cases. To investigate the impact of feature correlation on these differences, we calculated the maximum correlation (MC) between one feature and any other feature

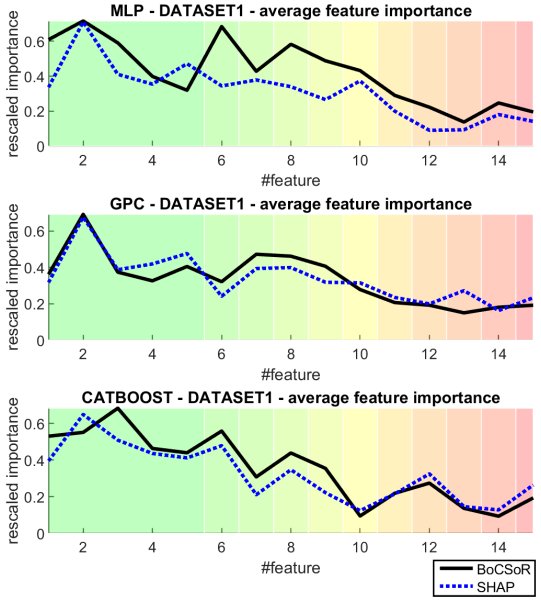

Figure 2: Experimental results with Dataset1 and three ML approaches. Feature importance computed via BoCSoR and SHAP with 10 repetitions and averaged by feature. The color of the background indicates the trend of the ground truth feature importance (from green to red, from high to low).

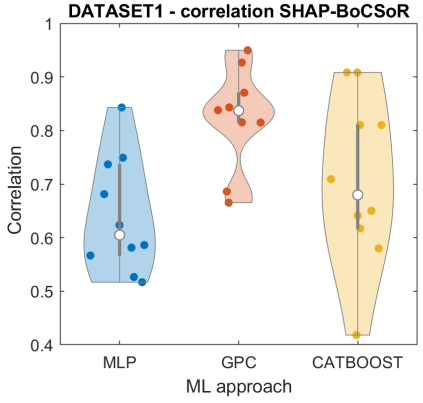

Figure 3: Correlation between feature importance computed via BoCSoR and SHAP with Dataset1 and three ML approaches.

in the dataset. Subsequently, we computed the MC for the five most similar and the five most dissimilar features in terms of importance for both BoCSoR and SHAP, and we rescaled the resulting values using a min-max procedure. The violin plots of the MC values obtained for each ML approach are presented in Figure 4. Our results demonstrate that the group of features where SHAP and BoCSoR differ the most is characterized by a higher MC, indicating a stronger correlation with other features in the dataset. This finding is consistent across all three ML approaches employed in our study.

For Dataset2, we computed the feature importance using both SHAP and BoCSoR, as shown in Figure 5. The solid and dotted lines in the figure represent the average feature

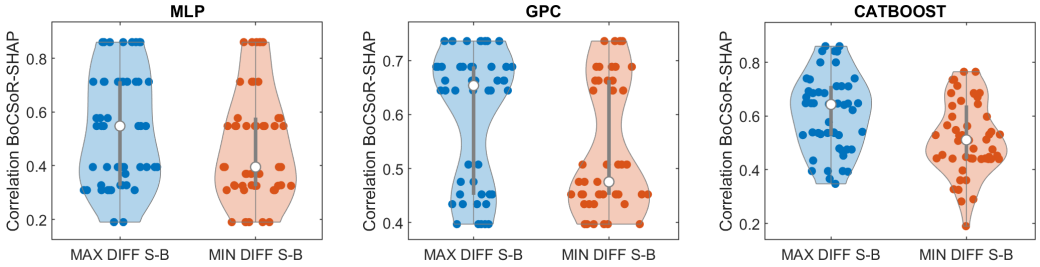

Figure 4: Experimental results on Dataset1 with three ML approaches. Violin plot of the maximum correlation between one feature and all the others considering the 5 features having the most similar normalized value for SHAP and BOCSOR, and the most dissimilar normalized value for SHAP and BOCSOR.

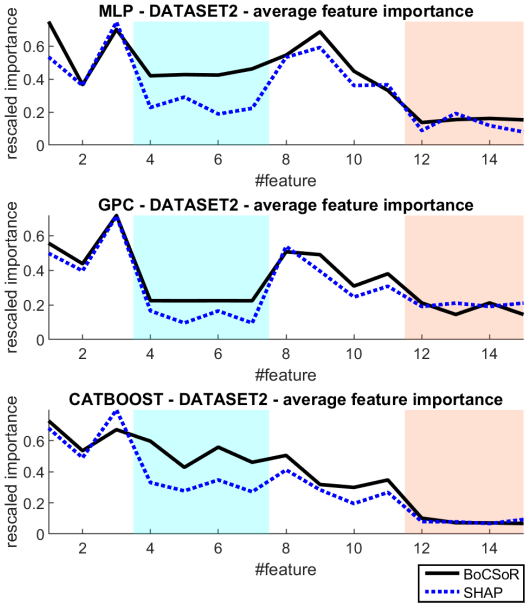

Figure 5: Experimental results with Dataset2 and three ML approaches. The feature importance computed via BoCSoR and SHAP with 10 repetitions and averaged by feature. The color of the background indicates the groups of highly correlated features (blue if important and replicated, orange if unimportant and replicated).

importance measure provided by BoCSoR and SHAP, respectively. We also added a colored background to the figure to highlight the groups of highly correlated features, with blue indicating important and replicated features, and orange indicating unimportant and replicated features. Compared to SHAP, BoCSoR exhibits greater importance values for informative features (blue background) that are replicated and therefore correlated. This behavior is consistent across all three ML approaches, suggesting the robustness of BoCSoR to feature correlation. Furthermore, the figure demonstrates that BoCSoR displays a more stable behavior while measuring the importance of informative and replicated features, except for Catboost. These findings confirm the consistency and effectiveness of BoCSoR in measuring feature importance for replicated and correlated features.

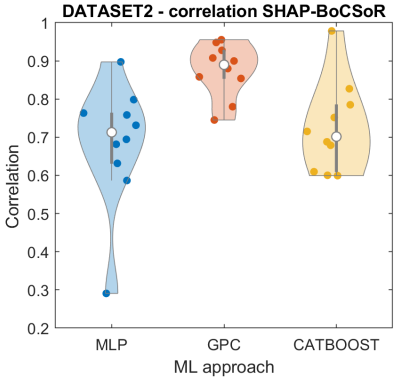

Figure 6: Correlation between the feature importance computed via BoCSoR and SHAP on Dataset2 with three ML approaches.

Table 2
Classification accuracy for Emotion and Social task of the HCP data. Average % ± standard deviation.

Task	GPC	MLP	Catboost
Emotion	64.36 ± 2.14	70.32 ± 3.72	73.29 ± 2.050
Social	62.14 ± 2.63	81.71 ± 1.90	80.05 ± 2.553

We assessed the agreement between SHAP and BoCSoR in Figure 6 and obtained results similar to those obtained for Dataset1.

Finally, we grouped replicated features based on their maximum and minimum importance and compared the difference between SHAP and BoCSoR for each group. Figure 7 shows that, regardless of the ML approach used, BoCSoR provides, on average, higher feature importance for important but highly correlated features (i.e., the median of their difference is > 0), and lower feature importance for less important but highly correlated features (i.e., the median of their difference is $<= 0$).

4.2. Real Datasets: fMRI data from HCP

To classify emotional states from the HCP data, we employed the three ML approaches introduced at the beginning of Section 4. Table 2 presents the accuracy scores obtained using a Monte Carlo 10 cross-fold validation schema. Catboost and MLP achieved greater average accuracy compared

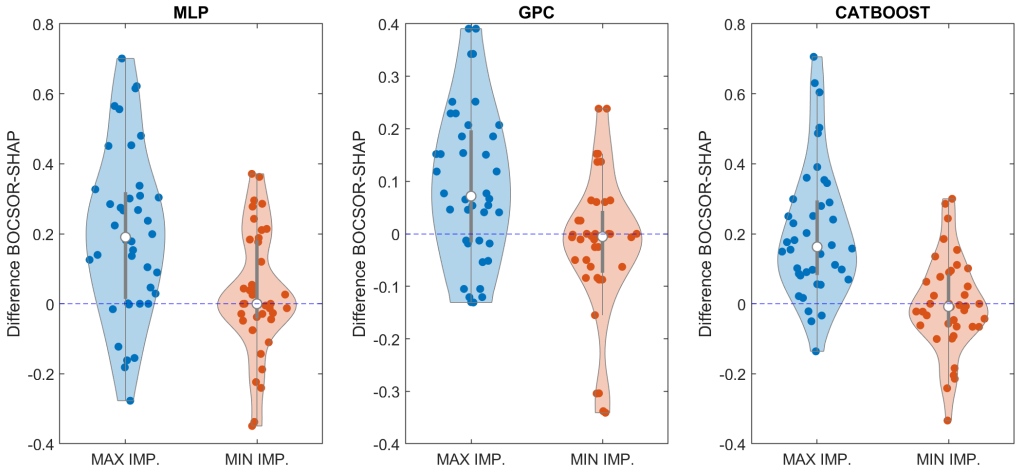

Figure 7: Experimental results on Dataset2 with three ML approaches. Violin plot of the difference between the feature importance measure computed via SHAP and BOCSOR. The difference is presented for the most and the less important features in the dataset when those are replicated (and so correlated).

Table 3
Percentage of of boundary instances from which BoCSoR can generate a counterfactual. Emotion and Social tasks with the HCP data. CatBoost classifier. Average % ± standard deviation.

Steps	Emotion	Social
2	53.45 ± 2.73	38.51 ± 4.65
3	64.29 ± 3.47	57.02 ± 3.41
5	76.31 ± 4.91	73.10 ± 3.96
10	87.98 ± 1.40	85.65 ± 3.40
20	93.51 ± 1.74	95.42 ± 1.80

to GPC, with an increase in accuracy of between 5 and 10 percent. When considering standard deviation, Catboost and MLP had roughly comparable accuracy for the Social task, while for the Emotion task, Catboost outperformed MLP in terms of accuracy. Thus, we employ Catboost for further experimentation on the HCP data, as overall it appears to be the most accurate.

The *steps* parameter in BoCSoR determines the granularity of the counterfactual search around the decision boundary. Opting for a low value for *steps* may result in faster counterfactual searches but poorer space exploration, and consequently, fewer counterfactuals. To investigate the effect of different parameterizations on the number of generated counterfactuals, we conducted experiments using *steps* values of 2, 3, 5, 10, and 20 with the HCP data. Table 3 presents the results obtained from these experiments.

Based on the results in Table 3, the percentage of instances from which BoCSoR can generate at least one counterfactual exhibits a sublinear increase as the *steps* parameter increases. For our subsequent experiments, a *steps* parameterization of 10 can be considered a good trade-off, as the improvement in the percentage of instances that result in the generation of a counterfactual appears to plateau beyond this value.

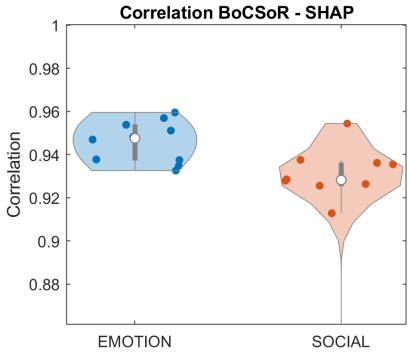

Figure 8: Experimental results on HCP-Emotion Processing. Violin plot of the correlation between BOCSOR and SHAP obtained via 10 repetitions with the 2 considered tasks. All the correlation values shown below correspond to p-values lower than 0.05.

We evaluated BoCSoR in terms of agreement with SHAP as measured through the Pearson correlation coefficient on the HCP data using Catboost as classifier. The results obtained are presented in Fig. 8 via violin plots.

Our findings show that BoCSoR and SHAP are significantly and positively correlated, but partially disagree.

We tested if the behavior occurring with Dataset1 and Dataset2 was confirmed with the HCP data. Thus, we rescaled the values of BOCSOR and SHAP via a min-max procedure to make them comparable. Then, the five features characterized by the most similar feature importance values according to SHAP and BOCSOR were considered, together with the five most dissimilar ones. For each feature, the MC is computed. According to our results, BoCSoR and SHAP disagreed on the ranking of the most correlated features in the Emotion Processing task. The graphical representation of the corresponding MC values is presented in Fig. 9.

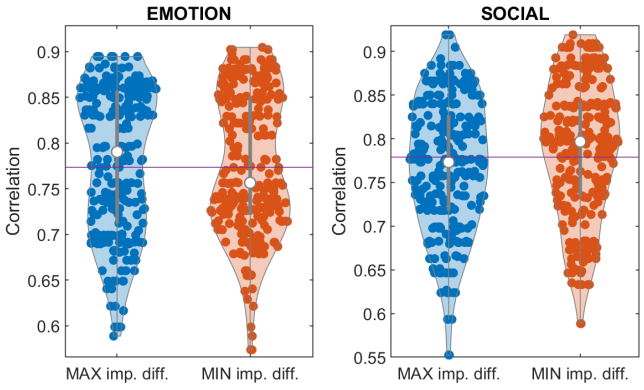

Figure 9: Experimental results on HCP-Emotion Processing. Violin plot of the maximum correlation between one feature and all the others considering the 5 features that have the most similar normalized value for SHAP and BOCSOR, and the 5 most dissimilar normalized values for SHAP and BOCSOR. The purple line indicates the average maximum correlation considering all the features in the data.

Our results show that there is a disagreement in the ranking of the most correlated features between BoCSoR and SHAP for the Emotion task data. Specifically, the 95% confidence interval of the maximum correlation for the features with the highest disagreement (agreement) between SHAP and BoCSoR is higher (smaller) than the average correlation of all features. In contrast, the data from the HCP-Social task exhibits partial overlap between the 95% confidence intervals of the maximum correlation for the features with the highest disagreement and agreement between SHAP and BoCSoR.

We further analyzed the importance matrices as adjacency structures, as shown in Figure 10 (left), and found that for the Emotion Processing task data, several regions of the temporal and occipital lobes have high importance levels. Previous research has established that visual cortices (LOcC and MVOcC) in the occipital lobe process visual information, while the temporal associative areas (ITG and FuG) from the fusiform gyrus are responsible for processing the emotional meaning of objects, particularly faces [58].

Similarly, for the HCP-Social task data, we observed consistent evidence in the functional connectivity of the evoked activity, as illustrated in Figure 10 (right). The fusiform gyrus appears to have a central role (highest node degree) in connecting with the cingulate gyrus (CG). The Basal Ganglia (BG) and the visual cortex (MVOcC) are linked, indicating the cooperation between the emotion and cognition regions and a visual area. Additionally, the BG is associated with a multisensory association cortex (STG), and regions dedicated to attention and visuospatial perception (SPL) and audiovisual emotional processing (MTG) are also linked. Furthermore, a visual cortex (MVOcC) and the somatosensory cortex, which represents body sensation information (PoG), also have high importance values.

5. Discussion

This study presents a novel XAI methodology, BoCSoR, which is employed to elucidate the classifications provided by various ML approaches on synthetic and fMRI benchmark datasets. BoCSoR offers a global measure of feature importance based on local counterfactuals, showcasing its reliability, resilience to feature correlation, and computational efficiency in comparison to other state-of-the-art feature importance measures such as SHAP. Overall, this work contributes to the advancement of the XAI field by providing a new tool for understanding and interpreting the decisions made by ML models, which can have significant implications across a broad range of applications. In this study, BoCSoR explained the classifications provided by three distinct ML approaches (CatBoost, Multilayer Perceptron, and Gaussian Process Classifier) on two synthetic datasets and two fMRI benchmark datasets. By utilizing only local counterfactuals, BoCSoR offers a global measure of feature importance. Our experiments demonstrate that it is a reliable and robust method for feature importance assessment, even when faced with correlated features. In comparison to other state-of-the-art feature importance measures, such as SHAP, BoCSoR is considerably more computationally efficient.

These properties are of importance for all XAI applications, especially for those addressing medical decision support systems [29]. Indeed, given the wide variety of features that can be extracted from the same physiological measures and the absence of a gold standard, it is difficult to determine the optimal subset of features to consider for a given analysis. To avoid losing information, many different features are often considered [59], which are highly likely to have some degree of mutual correlation, possibly hampering the reliability of SHAP and similar approaches [60; 61].

BoCSoR employs a model-agnostic, instance-based, and exogenous XAI counterfactual approach. This means it can be used to explain any classification approach for tabular data. At the same time, BoCSoR does provide global feature importance as an explanation form. Thus, BoCSoR belongs to a novel thread in the XAI literature, in which different explanation strategies are combined to generate new and better ones [41].

If compared to other approaches able to combine feature importance and counterfactual explanations, BoCSoR (i) provides global feature importance [37; 39; 38], (ii) can handle data with hundreds of features [39], and (iii) does not rely on predetermined structured knowledge [36].

The presented experimental activities was aimed at answering three research questions.

First, *can BoCSoR be considered reliable as a measure of feature importance?* The relative nature of the ground truth for feature importance allows us to make qualitative considerations, thus validating the trend of the feature importance measure presented in this study. As seen in Fig. 2, BoCSoR displays fewer abrupt drops in importance than SHAP, particularly with GPC and Catboost, which is expected given that noise within the data is incrementally added from the $6th$ to the $15th$ feature. Lastly, assuming SHAP as the

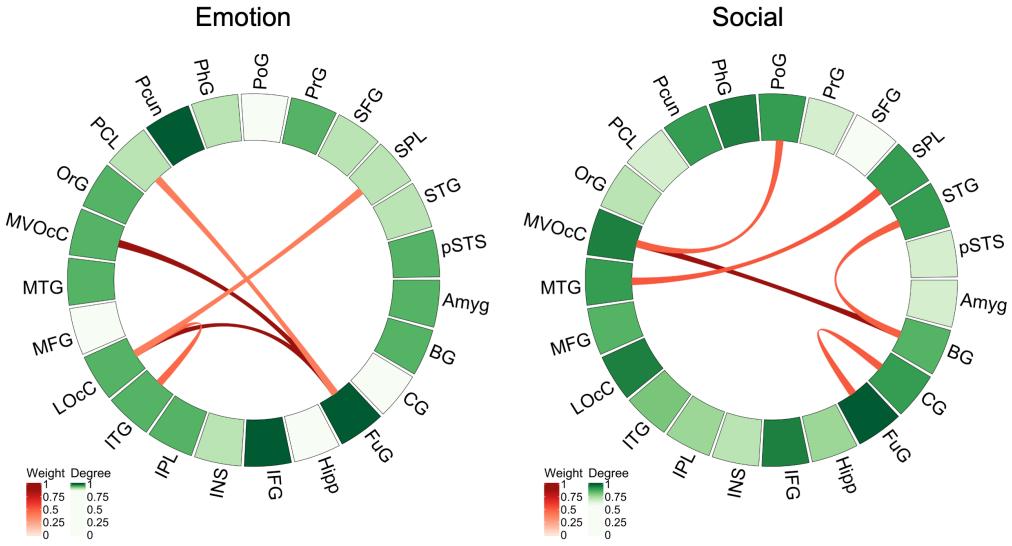

Figure 10: Circular plots expressing the functional connections between macroregions in the Emotion (Left) and Social (Right) HCP tasks. Green shade indicates the node degree and red shades represent the connection strength. In order to emphasize the most relevant functional connections, just the 5 top-most important are displayed.

baseline measure, the correlation between SHAP and the proposed measure (Fig. 3, 6 and 8), verifies that BoCSoR effectively captures feature importance across all the ML approaches utilized.

Second, *does BoCSoR address the main weaknesses of SHAP? That is, does it offer lower computational cost and less sensitivity to features' correlation?* From our experimental results, BoCSoR results in a computational cost that is an order of magnitude smaller as compared to SHAP, for both synthetic and HCP data. Furthermore, it is apparent in all the analyses that the difference between feature importance calculated using BoCSoR and SHAP is related to the correlation between the features. Features exhibiting high correlation with others correspond to the largest differences in SHAP vs BoCSoR feature importance (see Fig. 4, 7 and 9). This trend remains consistent for the two synthetic datasets, irrespective of the ML approach employed. Yet, this behavior does not occur with the HCP data for the Social task, which is also the task with which the model performs the worst. BoCSoR's robustness to feature correlation is also evident from a qualitative analysis of the results displayed in Fig. 5. For all three ML approaches considered, BoCSoR yields a higher average importance measure for informative features (blue background) despite being replicated (and thus correlated), and this behavior is consistent across all three ML approaches. On the other hand, for the group of replicated non-informative features (orange background), the difference between BoCSoR and SHAP is not qualitatively apparent but is still quantitatively confirmed by the results in Fig. 7. Lastly, except for Catboost, BoCSoR exhibits less fluctuating behavior while measuring the importance of informative and replicated features, which substantiates its consistency.

Third, *can BoCSoR be employed to extract knowledge from the trained AI model?* The prominent connections highlighted by our method seem coherent with the most salient functional connections reported in the literature for the HCP emotional [40], [62], [63] and social tasks [40], [64], [65]. Indeed, the emotional task, built over a sequence of contrasts between human faces expressing strong emotion (fear, panic, anger, etc.) and simple emotionless object shapes, is known to recruit both cortical visual areas and face emotion recognition regions, in accordance with the connections we found with BoCSoR. Similarly, the social task evokes both motor, somatosensory, and associative cortical regions and these were highlighted by BoCSoR importance rating. Therefore, BoCSoR appears to identify the most salient functional connections which are expected to be most active in the HCP cognitive tasks analyzed in this paper.

6. Conclusion

We have introduced a new measure of global feature importance, namely BoCSoR. BoCSoR utilizes local counterfactuals obtained from instances close to the decision boundary of a classifier to determine which features, if modified, are most likely to result in a change of classification. Our experiments show that BoCSoR outperforms SHAP, which is considered the gold standard for feature importance in the literature. BoCSoR is more reliable, less sensitive to feature correlation, and less computationally expensive.

The robustness of BoCSoR to the correlation among features makes it particularly suitable for the analysis of physiological data, where a high degree of correlation is expected between multi-domain signals collected from the

same subject. This is especially relevant in partial (i.e., ROI-wise) brain data, making our approach an excellent candidate for any whole-brain neuroimaging or neuromonitoring study.

To maintain low computational complexity (i.e., an order of magnitude less wall-clock time compared to SHAP), the counterfactual search utilized in BoCSoR is based on a linear search starting from the neighborhood of instances close to the decision boundary. However, this approach does not guarantee the minimal distance between the instance and the obtained counterfactual, nor does it guarantee the best approximation of the decision boundary.

On the other hand, modifying the approach for generating counterfactuals could influence the feature importance measurement. Given the way BoCSoR generates feature importance, two properties of counterfactuals can have a substantial impact: similarity and sparsity [31]. The similarity of a counterfactual approach ensures the smallest possible distance between an instance and its counterfactual [31]. As demonstrated in our experimentation, a more fine-grained search around the decision boundary can guarantee an adequate number of counterfactuals to derive the feature importance. The sparsity of a counterfactual approach ensures that there is the lowest number of modified features between an instance and its counterfactual [31]. If employed by BoCSoR, this may result in fewer relevant features per decision boundary instance (Algorithm 2), and consequently, a feature importance measure skewed towards a few features. However, according to experimentation in [31], approaches that offer better similarity, such as CBCE [66], and better sparsity, like DiCE [33], are up to three orders of magnitude more computationally expensive than brute-force approaches (i.e., the one used in BoCSoR). Nonetheless, the growing literature on counterfactual explanation procedures provides more sophisticated approaches that can strike a better balance between decision boundary approximation and computational cost. Future research will explore these directions.

In summary, the proposed method, BoCSoR, offers an efficient and reliable means of identifying the most important features for classification in the context of physiological data analysis. We believe that this approach will prove its usefulness in various neuroimaging studies, where the identification of critical features is crucial for the development of accurate and interpretable diagnostic tools.

Declaration of Competing Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

Acknowledgments

Work partially supported by Horizon 2020 Program under GA 101017727 of the project “EXPERIENCE”; and Italian Ministry of University and Research (MUR) in the frameworks: "FAIR" PE00000013 Spoke1 "Human-centered AI"; National Center for Sustainable Mobility MOST/Spoke10; "Reasoning" project, PRIN 2020 LS Programme, Project number 2493 04-11-2021; "THE" cod. ECS00000017 - CUP I53C22000780001; and FoReLab project (Departments of Excellence).

References

[1] Sarma GP, Reinertsen E, Aguirre A, Anderson C, Batra P, Choi SH, et al. Physiology as a Lingua Franca for clinical machine learning. Patterns. 2020;1(2):100017.

[2] Başar E, Bullock TH. Brain dynamics: Progress and perspectives. 2012.

[3] Kiani M, Andreu-Perez J, Hagras H, Rigato S, Filippetti ML. Towards Understanding Human Functional Brain Development with Explainable Artificial Intelligence: Challenges and Perspectives. IEEE Computational Intelligence Magazine. 2022;17(1):16-33.

[4] Amann J, Vetter D, Blomberg SN, Christensen HC, Coffee M, Gerke S, et al. To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. PLOS Digital Health. 2022;1(2):e0000016.

[5] Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ. XAI—Explainable artificial intelligence. Science robotics. 2019;4(37):eaay7120.

[6] Angelov PP, Soares EA, Jiang R, Arnold NI, Atkinson PM. Explainable artificial intelligence: an analytical review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2021;11(5):e1424.

[7] Vu MAT, Adalı T, Ba D, Buzsáki G, Carlson D, Heller K, et al. A shared vision for machine learning in neuroscience. Journal of Neuroscience. 2018;38(7):1601-7.

[8] Langlotz CP, Allen B, Erickson BJ, Kalpathy-Cramer J, Bigelow K, Cook TS, et al. A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 NIH/RSNA/ACR/The Academy Workshop. Radiology. 2019;291(3):781-91.

[9] Fellous JM, Sapiro G, Rossi A, Mayberg H, Ferrante M. Explainable artificial intelligence for neuroscience: behavioral neurostimulation. Frontiers in neuroscience. 2019;13:1346.

[10] Kriegeskorte N, Mur M, Bandettini PA. Representational similarity analysis-connecting the branches of systems neuroscience. Frontiers in systems neuroscience. 2008:4.

[11] Morabito FC, Ieracitano C, Mammone N. An explainable Artificial Intelligence approach to study MCI to AD conversion via HD-EEG processing. Clinical EEG and Neuroscience. 2023;54(1):51-60.

[12] Islam MS, Hussain I, Rahman MM, Park SJ, Hossain MA. Explainable Artificial Intelligence Model for Stroke Prediction Using EEG Signal. Sensors. 2022;22(24):9859.

[13] Karpov OE, Grubov VV, Maksimenko VA, Kurkin SA, Smirnov NM, Utyashev NP, et al. Extreme value theory inspires explainable machine learning approach for seizure detection. Scientific Reports. 2022;12(1):11474.

[14] Galazzo IB, Cruciani F, Brusini L, Salih A, Radeva P, Storti SF, et al. Explainable artificial intelligence for magnetic resonance imaging aging brainprints: Grounds and challenges. IEEE Signal Processing Magazine. 2022;39(2):99-116.

[15] Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human–computer collaboration for skin cancer recognition. Nature Medicine. 2020;26(8):1229-34.

[16] Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: contextualizing explainable machine learning for clinical end use. In: Machine learning for healthcare conference. PMLR; 2019. p. 359-80.

[17] Foulsham M, Hitchen B, Denley A. GDPR: how to achieve and maintain compliance. 2019.

[18] Schoenborn JM, Althoff KD. Recent Trends in XAI: A Broad Overview on current Approaches, Methodologies and Interactions.

In: ICCBR Workshops; 2019. p. 51-60.

[19] Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence. 2019;267:1-38.

[20] van der Waa J, Nieuwburg E, Cremers A, Neerincx M. Evaluating XAI: A comparison of rule-based and example-based explanations. Artificial Intelligence. 2021;291:103404.

[21] Delaney E, Greene D, Keane MT. Instance-based counterfactual explanations for time series classification. In: International Conference on Case-Based Reasoning. Springer; 2021. p. 32-47.

[22] Afchar D, Guigue V, Hennequin R. Towards rigorous interpretations: a formalisation of feature attribution. In: International Conference on Machine Learning. PMLR; 2021. p. 76-86.

[23] Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. Journal of Biomedical Informatics. 2021;113:103655.

[24] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017;30.

[25] Mosca E, Szigeti F, Tragianni S, Gallagher D, Groh G. SHAP-Based Explanation Methods: A Review for NLP Interpretability. In: Proceedings of the 29th International Conference on Computational Linguistics; 2022. p. 4593-603.

[26] Pat N, Wang Y, Bartonicek A, Candia J, Stringaris A. Explainable Machine Learning Approach to Predict and Explain the Relationship between Task-based fMRI and Individual Differences in Cognition. bioRxiv. 2022:2020-10.

[27] Kumar IE, Venkatasubramanian S, Scheidegger C, Friedler S. Problems with Shapley-value-based explanations as feature importance measures. In: International Conference on Machine Learning. PMLR; 2020. p. 5491-500.

[28] Marcílio WE, Eler DM. From explanations to feature selection: assessing shap values as feature selection mechanism. In: 2020 33rd SIBGRAPI conference on Graphics, Patterns and Images (SIBGRAPI). Ieee; 2020. p. 340-7.

[29] Dai Y, Niu L, Wei L, Tang J. Feature Selection in High Dimensional Biomedical Data Based on BF-SFLA. Frontiers in Neuroscience. 2022;16.

[30] Wiratunga N, Wijekoon A, Nkisi-Orji I, Martin K, Palihawadana C, Corsar D. Actionable feature discovery in counterfactuals using feature relevance explainers. CEUR Workshop Proceedings; 2021. .

[31] Guidotti R. Counterfactual explanations and how to find them: literature review and benchmarking. Data Mining and Knowledge Discovery. 2022:1-55.

[32] Sokol K, Hepburn A, Poyiadzi R, Clifford M, Santos-Rodriguez R, Flach P. FAT Forensics: A Python Toolbox for Implementing and Deploying Fairness, Accountability and Transparency Algorithms in Predictive Systems. Journal of Open Source Software. 2020;5(49):1904. Available from: https://doi.org/10.21105/joss.01904.

[33] Mothilal RK, Sharma A, Tan C. Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 conference on fairness, accountability, and transparency; 2020. p. 607-17.

[34] Stepin I, Alonso JM, Catala A, Pereira-Fariña M. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. IEEE Access. 2021;9:11974-2001.

[35] Setzu M, Guidotti R, Monreale A, Turini F, Pedreschi D, Giannotti F. Glocalx-from local to global explanations of black box AI models. Artificial Intelligence. 2021;294:103457.

[36] Galhotra S, Pradhan R, Salimi B. Feature attribution and recourse via probabilistic contrastive counterfactuals. In: Proceedings of the ICML Workshop on Algorithmic Recourse; 2021. p. 1-6.

[37] Vlassopoulos G, van Erven T, Brighton H, Menkovski V. Explaining predictions by approximating the local decision boundary. arXiv preprint arXiv:200607985. 2020.

[38] Ribeiro MT, Singh S, Guestrin C. Anchors: High-precision model-agnostic explanations. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32; 2018. .

[39] Laugel T, Renard X, Lesot MJ, Marsala C, Detyniecki M. Defining Locality for Surrogates in Post-hoc Interpretablity. In: Workshop on Human Interpretability for Machine Learning (WHI)-International Conference on Machine Learning (ICML); 2018. .

[40] Barch DM, Burgess GC, Harms MP, Petersen SE, Schlaggar BL, Corbetta M, et al. Function in the human connectome: task-fMRI and individual differences in behavior. NeuroImage. 2013 10;80:169-89. Available from: https://pubmed.ncbi.nlm.nih.gov/23684877/.

[41] Kommiya Mothilal R, Mahajan D, Tan C, Sharma A. Towards unifying feature attribution and counterfactual explanations: Different means to the same end. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society; 2021. p. 652-63.

[42] Barr B, Xu K, Silva C, Bertini E, Reilly R, Bruss CB, et al. Towards ground truth explainability on tabular data. arXiv preprint arXiv:200710532. 2020.

[43] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825-30.

[44] Guyon I. Design of experiments of the NIPS 2003 variable selection benchmark. In: NIPS 2003 workshop on feature extraction and feature selection. vol. 253; 2003. p. 40.

[45] Yang M, Kim B. Benchmarking attribution methods with relative feature importance. arXiv preprint arXiv:190709701. 2019.

[46] Fan L, Li H, Zhuo J, Zhang Y, Wang J, Chen L, et al. The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. Cerebral Cortex. 2016;26:3508-26. Available from: http://dx.doi.org/10.1093/cercor/bhw157.

[47] Hariri AR, Tessitore A, Mattay VS, Fera F, Weinberger DR. The amygdala response to emotional stimuli: a comparison of faces and scenes. Neuroimage. 2002;17(1):317-23.

[48] Castelli F, Frith C, Happé F, Frith U. Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. Brain. 2002;125(8):1839-49.

[49] Frolov N, Kabir MS, Maksimenko V, Hramov A. Machine learning evaluates changes in functional connectivity under a prolonged cognitive load. Chaos: An Interdisciplinary Journal of Nonlinear Science. 2021;31(10):101106.

[50] Rodriguez CI, Vergara VM, Davies S, Calhoun VD, Savage DD, Hamilton DA. Detection of prenatal alcohol exposure using machine learning classification of resting-state functional network connectivity data. Alcohol. 2021;93:25-34.

[51] Ji Y, Yang C, Liang Y, et al. A Multiview Deep Learning Method for Brain Functional Connectivity Classification. Computational Intelligence and Neuroscience. 2022;2022.

[52] Cao J, Garro EM, Zhao Y. EEG/fNIRS Based Workload Classification Using Functional Brain Connectivity and Machine Learning. Sensors. 2022;22(19):7623.

[53] Jie B, Shen D, Zhang D. Brain connectivity hyper-network for MCI classification. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part II 17. Springer; 2014. p. 724-32.

[54] Du Y, Fu Z, Calhoun VD. Classification and prediction of brain disorders using functional connectivity: promising but challenging. Frontiers in neuroscience. 2018;12:525.

[55] Sendi MS, Chun JY, Calhoun VD. Visualizing functional network connectivity difference between middle adult and older subjects using an explainable machine-learning method. In: 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE). IEEE; 2020. p. 955-60.

[56] Ranjan A, Singh VP, Singh AK, Thakur AK, Mishra RB. Classifying Brain State in Sentence Polarity Exposure: An ANN Model for fMRI Data. Revue d'Intelligence Artificielle. 2020;34(3):361-8.

[57] Challis E, Hurley P, Serra L, Bozzali M, Oliver S, Cercignani M. Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI. NeuroImage. 2015;112:232-43.

[58] Kanwisher N, Yovel G. The fusiform face area: a cortical region specialized for the perception of faces. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2006 12;361:2109-28. Available from: https://pubmed.ncbi.nlm.nih.gov/17118927/.

[59] Polat K, Güneş S. A hybrid approach to medical decision support systems: Combining feature selection, fuzzy weighted pre-processing and AIRS. Computer methods and programs in biomedicine. 2007;88(2):164-74.

[60] Nohara Y, Matsumoto K, Soejima H, Nakashima N. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. Computer Methods and Programs in Biomedicine. 2022;214:106584.

[61] Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, Acharya UR. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). Computer Methods and Programs in Biomedicine. 2022:107161.

[62] Markett S, Jawinski P, Kirsch P, Gerchen MF. Specific and segregated changes to the functional connectome evoked by the processing of emotional faces: A task-based connectome study. Scientific Reports 2020 10:1. 2020 3;10:1-14. Available from: https://www.nature.com/articles/s41598-020-61522-0.

[63] Weathersby FL, King JB, Fox JC, Loret A, Anderson JS. Functional connectivity of emotional well-being: Overconnectivity between default and attentional networks is associated with attitudes of anger and aggression. Psychiatry research Neuroimaging. 2019 9;291:52. Available from: /pmc/articles/PMC6708495//pmc/articles/PMC6708495/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6708495/.

[64] Marchetti A, Baglio F, Costantini I, Dipasquale O, Savazzi F, Nemni R, et al. Theory of mind and the whole brain functional connectivity: Behavioral and neural evidences with the Amsterdam Resting State Questionnaire. Frontiers in Psychology. 2015 12;6:1855.

[65] Ilzarbe D, de la Serna E, Baeza I, Rosa M, Puig O, Calvo A, et al. The relationship between performance in a theory of mind task and intrinsic functional connectivity in youth with early onset psychosis. Developmental cognitive neuroscience. 2019 12;40. Available from: https://pubmed.ncbi.nlm.nih.gov/31791005/.

[66] Keane MT, Smyth B. Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable AI (XAI). In: Case-Based Reasoning Research and Development: 28th International Conference, ICCBR 2020, Salamanca, Spain, June 8–12, 2020, Proceedings 28. Springer; 2020. p. 163-78.