Efficient Interpretation of Deep Learning Models Using Graph Structure and Cooperative Game Theory: Application to ASD Biomarker Discovery

Xiaoxiao Li*[1], Nicha C. Dvornek[2], Yuan Zhou[1], Juntang Zhuang[1],
Pamela Ventola[3], and James S. Duncan[1,2,4,5]

[1] Biomedical Engineering, Yale University, New Haven, CT USA
[2] Radiology & Biomedical Imaging, Yale School of Medicine, New Haven, CT USA
[3] Child Study Center, Yale School of Medicine, New Haven, CT USA
[4] Electrical Engineering, Yale University, New Haven, CT, USA
[5] Statistics & Data Science Yale University New Haven, CT, USA

Abstract. Discovering imaging biomarkers for autism spectrum disorder (ASD) is critical to help explain ASD and predict or monitor treatment outcomes. Toward this end, deep learning classifiers have recently been used for identifying ASD from functional magnetic resonance imaging (fMRI) with higher accuracy than traditional learning strategies. However, a key challenge with deep learning models is understanding just what image features the network is using, which can in turn be used to define the biomarkers. Current methods extract biomarkers, i.e., important features, by looking at how the prediction changes if "ignoring" one feature at a time. However, this can lead to serious errors if the features are conditionally dependent. In this work, we go beyond looking at only individual features by using Shapley value explanation (SVE) from cooperative game theory. Cooperative game theory is advantageous here because it directly considers the interaction between features and can be applied to any machine learning method, making it a novel, more accurate way of determining instance-wise biomarker importance from deep learning models. A barrier to using SVE is its computational complexity: 2^N given N features. We explicitly reduce the complexity of SVE computation by two approaches based on the underlying graph structure of the input data: 1) only consider the centralized coalition of each feature; 2) a hierarchical pipeline which first clusters features into small communities, then applies SVE in each community. Monte Carlo approximation can be used for large permutation sets. We first validate our methods on the MNIST dataset and compare to human perception. Next, to insure plausibility of our biomarker results, we train a Random Forest (RF) to classify ASD/control subjects from fMRI and compare SVE results to standard RF-based feature importance. Finally, we show initial results on ranked fMRI biomarkers using SVE on a deep learning classifier for the ASD/control dataset.

[1]* To whom correspondence should be addressed, `xiaoxiao.li@yale.edu`

1 Introduction

Autism spectrum disorder (ASD) affects the structure and function of the brain. To better target the underlying roots of ASD for diagnosis and treatment, efforts to identify reliable biomarkers are growing [1]. Deep learning models have been used in fMRI analysis [2], which is used to characterize the brain changes that occur in ASD [3]. However, how the different brain regions coordinate on the deep convolutional neural network (DNN) has not been previously explored. When features are not independent, Shapley value explanation (SVE) is a useful tool to study each feature's contribution [4,5,6]. The methods are based on fundamental concepts from cooperative game theory [7], which assigns a unique distribution (among the players) of a total surplus generated by the coalition of all players in the cooperative game. However, if the interactive features' dimensions are high, SVE becomes computationally consuming (exponential time complexity).

The innovations of this study include: 1) We applied SVE on interactive features' prediction power analysis; 2) Our proposed method does not require retraining the classifier; 3) To handle the high dimensional inputs of the DNN classifier, we propose two methods to reduce the dimension of SVE testing features, once the underlying graph structure of features is defined; and 4) Different from kernel SHAP proposed in [4], as a model interpreter, our proposed methods do not require model approximation. In section 2, we introduce the background on cooperative game theory. In section 3, we propose the two approaches to approximate Shapley value. We also show the approximation is true under certain assumptions. Three experiments are given in section 4 to show the feasibility and advantage of our proposed methods.

2 Background on Cooperative Game Theory

2.1 Shapley Value

Our approach to analyzing the contributions of individual nodes to the overall network is the assignment of Shapley values. The Shapley value is a means of fairly portioning the collective profit attained by a coalition of players, based on the relative contributions of the players in some game. Let $\mathcal{N} = \{1, 2, \ldots, N\}$ be the set of all the players, $S \subset \mathcal{N}$ be a subset of players forming a coalition within this game, and $v : 2^{\mathcal{N}} \to \mathbb{R}$ be the function that assigns a real numbered profit to the subset S of players. By definition, for any v, $v(\varnothing) = 0$, here $\varnothing$ is the empty set. A Shapley value is assigned by a Shapley function $\Phi : \mathcal{N} \to \mathbb{R}$, which associates each player in $\mathcal{N}$ with a real number and which is uniquely defined by the following axioms [7]: 1. *Efficiency*; 2. *Symmetry*; 3. *Dummy*; and 4. *Additivity*. In our context, we are interested in the brain regions that discriminate ASD and control subjects. Classification prediction score is the total value to be distributed, and each brain region is a player, which will be assigned a unique reward (i.e. importance score) by its contribution to the classifier.

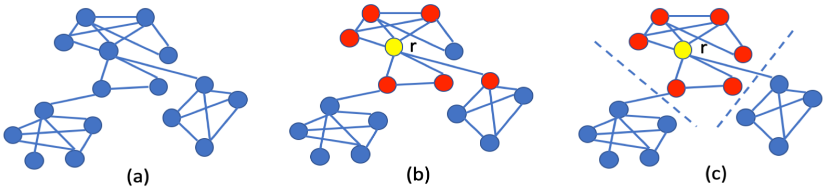


Fig. 1: a) Toy visualization of graph structure of the input data. When estimating the contribution of feature r (yellow), b) C-SVE considers r's directly connected neighbors (red) and c) H-SVE considers the community (red) to which r belongs.

2.2 Challenges Of Using Shapley Value

While Shapley values give a more accurate interpretation of the importance of each player in a coalition, their calculation is expensive. When the number of features (i.e., players in the game) is a massive N, the computational complexity is 2^N, which is especially expensive if the model is slow to run. We propose addressing this computational challenge by utilizing the graph structure of the data. Consider the case when the underlying graph structure of the data is sparsely connected, e.g., the brain functional network. Under this observation, we propose two approaches (Fig. 1) to simplify Shapley value calculation. *Method I* only considers the centralized coalition of each player to reduce the number of permutation cases by assigning weight 0 to features that rarely collaborate. *Method II* first applies community detection on the feature connectivity network to cluster similar features (forming different games and teams), then within the selected communities, assigns a feature's contribution by SVE.

3 Methods

In classification tasks, only certain features in a given input provide evidence for the classification decision. For a given prediction, the classifier assigns a relevance value to each input feature with respect to a class label $Y \in \mathcal{C}$. The probability of class Y for input $\boldsymbol{X} = (X_1, X_2, \dots, X_N)$ is given by the predictive score of the DNN model $f : \mathcal{D} \to \mathbb{R}^{|\mathcal{C}|}$ where $\mathcal{D}$ is the domain for input X and each component of the output of f represents the conditional probability of assigning a class label, i.e. $p(Y|\boldsymbol{X})$.

The basic idea used in prediction difference analysis [8] is that the relevance of a feature x_i can be estimated by measuring how the prediction changes if the feature is unknown. Here we extend this setting by considering the interaction of a set of different features instead of examining the features one by one. Denote the image corrupted at a feature set $S \subseteq \mathcal{N}$ as $\boldsymbol{X}_{\mathcal{N}\setminus S}$. To calculate $p(Y|\boldsymbol{X}_{\mathcal{N}\setminus S})$, following [8], we marginalize out the corrupted feature set S:

$$p(Y|\boldsymbol{X}_{\mathcal{N}\setminus S}) = \mathbb{E}_{\boldsymbol{X}_S \sim p(\boldsymbol{X}_S|\boldsymbol{X}_{\mathcal{N}\setminus S})} p(Y|\boldsymbol{X}_{\mathcal{N}\setminus S}, \boldsymbol{X}_S). \quad (1)$$

Denote $v_{\boldsymbol{X}}$ the importance score evaluation function for input $\boldsymbol{X}$. The prediction power for the rth feature is the weighted sum of all possible marginal contributions:

$$\varPhi_r(v_{\boldsymbol{X}}) = \frac{1}{|\mathcal{N}|} \sum_{S \subseteq \mathcal{N} \setminus \{r\}} \binom{|\mathcal{N}|-1}{|S|}^{-1} (v_{\boldsymbol{X}}(S \cup \{r\}) - v_{\boldsymbol{X}}(S)). \tag{2}$$

Similar to [5], we introduce the *importance score* of a feature set S

$$v_{\boldsymbol{X}}(S) \coloneqq \mathbb{E}_Y[-\log \frac{1}{p(Y|\boldsymbol{X})}|\boldsymbol{X}] - \mathbb{E}_Y[-\log \frac{1}{p(Y|\boldsymbol{X}_{\mathcal{N}\setminus S})}|\boldsymbol{X}], \tag{3}$$

which can be interpreted as the negative of the expected number of bits required to encode the output of the model based on the input $\boldsymbol{X}_{\mathcal{N}\setminus S}$.

Theorem 1 $\langle \mathcal{N} = \{1, 2, \ldots, N\}, v_{\boldsymbol{X}} \rangle$ *is a cooperative form game and* $\varPhi(v_{\boldsymbol{X}}) = (\varPhi_1, \varPhi_2, \ldots, \varPhi_N)$ *corresponds to the game's Shapley value.*

The proof can be directly borrowed from [6] showing it has a unique solution and satisfies Axioms 1-4.

An illustrative example is Boolean logic expression, $\mathrm{OR}((x_1, x_2)) = 1$ when x_1 or x_2 is one and zero otherwise for $\mathcal{N} = \{1, 2\}$, $\mathcal{D} = \{0, 1\} \times \{0, 1\}$. Suppose $p(Y = 1|\boldsymbol{X}) = \mathrm{OR}(\boldsymbol{X})$ and the base of the logarithm is 2. We aim to find the contributions of predicting 1 given input $\boldsymbol{X} = (1, 1)$. If both values of $\boldsymbol{X}$ are unknown, one can predict that the probability of the result being 1 is $\frac{3}{4}$. We have $v_{\boldsymbol{X}}(\varnothing) = 0 - (-\log \frac{1}{1}) = 0$, $v_{\boldsymbol{X}}(\{1\}) = v_{\boldsymbol{X}}(\{2\}) = 0 - (-\log(\frac{1}{1})) = 0$ and total value $v_{\boldsymbol{X}}(\{1, 2\}) = 0 - (-\log \frac{4}{3}) = \log \frac{4}{3}$. Therefore the contributions of each feature are: $\varPhi_1 = \frac{1}{2}[(v_{\boldsymbol{X}}(\{1\}) - v_{\boldsymbol{X}}(\varnothing)) + (v_{\boldsymbol{X}}(\{1, 2\}) - v_{\boldsymbol{X}}(\{2\}))] = \frac{1}{2}[(0 - 0) + (\log \frac{4}{3} - 0)] = \frac{1}{2} \log \frac{4}{3}$ and $\varPhi_2 = \frac{1}{2}[(v_{\boldsymbol{X}}(\{2\}) - v_{\boldsymbol{X}}(\varnothing)) + (v_{\boldsymbol{X}}(\{1, 2\}) - v_{\boldsymbol{X}}(\{1\}))] = \frac{1}{2}[(0 - 0) + (\log \frac{4}{3} - 0)] = \frac{1}{2} \log \frac{4}{3}$. The generated contributions reveal that both features contribute the same amount towards the prediction being 1 given input $(1, 1)$. In addition, we can interpret there is coalition between the two players, since $v_{\boldsymbol{X}}(\{1, 2\}) > v_{\boldsymbol{X}}(\{1\}) + v_{\boldsymbol{X}}(\{2\})$. However, we will get the myopic conclusion that both features are unimportant by only ignoring a single feature, because given one feature $X_i = 1$, $p = 1$ is for sure.

With the underlying structure of data, we have prior knowledge that some features of the data set are barely connected; in other words, there is very likely no coalition between these features. We define a connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with nodes $\mathcal{V}$ and edges $\mathcal{E}$. Given an adjacency matrix $\mathcal{A} = (a_{ij})$ of the undirected graph $\mathcal{G}$ (for example, the Pearson correlation of mean time series of brain regions), we use a threshold th to binarize a_{ij}, i.e. $a^b_{ij} = 1$ when $a_{ij} > th$ and zero otherwise, resulting in a sparsely connected graph.

3.1 Method I: Centralized Shapley Value Explanation (C-SVE)

For a given feature i, its 1-step connected *neighborhood* is defined by the set $\mathcal{N}(i) \coloneqq \{j \in \mathcal{V} | a^b_{ij} = 1\}$. As an approximation, we propose Centralized Shapley Value Explanation (C-SVE), which only calculates the marginal contribution when a feature collaborates with its neighbors.

Definition 1. *Given classifier f and sample $\boldsymbol{X}$, the C-SVE assigns the prediction* *power on feature r by*

$$\hat{\Phi}_r^C(v_{\boldsymbol{X}}) = \frac{1}{|\mathcal{N}(r)|} \sum_{S \subseteq \mathcal{N}(r) \setminus \{r\}} \binom{|\mathcal{N}(r)| - 1}{|S|}^{-1} (v_{\boldsymbol{X}}(S \cup \{r\}) - v_{\boldsymbol{X}}(S)). \tag{4}$$

The coefficients in front of the marginal contributions is a weighted transformation of the original SVE form (in Eq. (2)), where instead of assigning each permutation the same weight, sets not belonging to the *neighborhood* were assigned 0 weight. In practice, we can reject the non-coalition permutations and average the Shapley values for the remaining terms.

Theorem 2 *We have $\hat{\Phi}_{\boldsymbol{X}}^C(r) = \Phi_{\boldsymbol{X}}(r)$ almost surely if we have $X_r \perp \boldsymbol{X}_{\mathcal{N} \setminus \mathcal{N}(r)} | \boldsymbol{X}_U$ and $X_r \perp \ \boldsymbol{X}_{\mathcal{N} \setminus \mathcal{N}(r)} | \boldsymbol{X}_U, Y$ for any $U \subset \mathcal{N}(r) \setminus \{r\}$.*

The proof is shown in Appendix A. It is important to show that our proposed approximation is a good one. We can easily check that for $k \notin \mathcal{N}(r)$, the angle between the average time series $\bar{X}_r$ in ROI r and $\bar{X}_k$ in ROI k satisfies $\cos(\bar{X}_r, \bar{X}_k) < \epsilon$, which corresponds to the small edge weight (~ 0) in the graph that we created using Pearson correlation. Therefore we assume that $X_k \perp X_r$.

3.2 Method II: Hierarchical Shapley Value Explanation (H-SVE)

In method II, we approximate the Shapley value by a hierarchical approach: 1) detect communities in the graph, then 2) apply SVE in each community individually.

Modularity-based community detection We use the same undirected graph architecture defined in *Method I*, but use *greedy modularity method* [9] to divide all the features into non-overlapping communities. Then the whole features sets can be expressed by a combination of non-overlapping communities $\mathcal{N} = A_1 \bigcup A_2 \bigcup \cdots \bigcup A_M$ and the features in one community only cooperate within the group, hence are independent to those in the different communities. Therefore we can define different Shapley value rules in the different communities, but the Shapley values are comparable within and across communities.

Shapley value of each feature in the community With the assumption that different communities of players do not play in a game (rarely connect), we assume the communities of features are independent. In order to compare the feature importance in the whole brain, firstly we define the Shapley value for feature subset S in community A_i as

$$v_{\boldsymbol{X}}(S) := \mathbb{E}_Y[-\log \frac{1}{p(Y|\boldsymbol{X}_{A_i})} | \boldsymbol{X}_{A_i}] - \mathbb{E}_Y[-\log \frac{1}{p(Y|\boldsymbol{X}_{A_i \setminus S})} | \boldsymbol{X}_{A_i}]. \tag{5}$$

Definition 2. *Suppose the features are clustered into $\mathcal{N} = A_1 \bigcup A_2 \bigcup \cdots \bigcup A_M$. The H-SVE assigns the prediction power of feature r in A_i by*

$$\hat{\Phi}_r^H(v_{\boldsymbol{X}}) = \frac{1}{|A_i|} \sum_{S \subseteq A_i \setminus \{r\}} \binom{|A_i| - 1}{|S|}^{-1} (v_{\boldsymbol{X}}(S \cup \{r\}) - v_{\boldsymbol{X}}(S)), \ . \tag{6}$$

Algorithm 1 Approximating the prediction power of rth feature's value Φ_r

```
    Input: $\boldsymbol{X}$, a given instance; $m$, number of samples; $v$, importance score function
$\Phi_r \leftarrow 0$
for $j = 1$ to $m$ do
    choose a random permutation of features $\mathcal{O} \in \pi(\mathcal{N}(r))$
    choose a random instance $\hat{\boldsymbol{X}}$ from the training dataset
    $v_1 \leftarrow v(\tau(\boldsymbol{X}, \hat{\boldsymbol{X}}, Pre^r(\mathcal{O} \bigcup \{r\})))$
    $v_2 \leftarrow v(\tau(\boldsymbol{X}, \hat{\boldsymbol{X}}, Pre^r(\mathcal{O})))$
    $\Phi_r \leftarrow \Phi_r + (v_1 - v_2)$
end for
$\Phi_r \leftarrow \frac{\Phi_r}{m}$
(where $\mathcal{N}(r)$ is the neighborhood of $r$ in C-SVE or community of $r$ in H-SVE)
```

Theorem 3. *When* $\boldsymbol{X}_{A_1} \perp \boldsymbol{X}_{A_2} \perp \cdots \perp \boldsymbol{X}_{A_M}$, *we have* $\hat{\Phi}^H_r(v_{\boldsymbol{X}}) = \Phi_r(v_{\boldsymbol{X}})$ *almost surely.*

The proof is similar to the proof for *Theorem 2*.

3.3 Monte Carlo Approximation For Large Neighborhood

Although we simplify SVE by C-SVE or H-SVE methods, computation may still be challenging. For example: 1) in C-SVE, feature node r to be analyzed is densely connected with the other nodes and 2) in H-SVE, there exists large communities. Based on the alternative formulation of the Shapley value (Eq. (7)), let $\pi(\mathcal{N})$ be the set of all ordered permutations of $\mathcal{N}$. Let $Pre^r(O)$ be the set of players which are predecessors of player r in the order $O \in \pi(\mathcal{N})$, we have

$$\Phi_r(v_{\boldsymbol{X}}) = \frac{1}{|\mathcal{N}|!} \sum_{O \in \pi(\mathcal{N})} (v_{\boldsymbol{X}}(Pre^r(O) \cup \{r\}) - v_{\boldsymbol{X}}(Pre^r(O))). \tag{7}$$

We use the following Monte Carlo (MC) algorithm to approximate equation (4) and (6). We define:

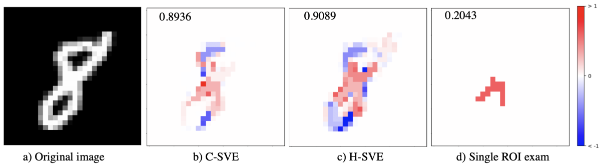
$$\tau(x, \hat{x}, S) = (z_1, z_2, \ldots, z_s), \quad z_i = \begin{cases} x_i;\ i \in S \\ \hat{x}_i;\ i \notin S \end{cases}. \tag{8}$$

Then the unbiased MC approximation can be expressed as in Algorithm 1. Given m, if $2^{|\mathcal{N}(r)|} \gg m$, we will apply MC approximation.

4 Experiments and Results

4.1 Validation on MNIST Dataset

In order to show the feasibility of the proposed two approaches, we test the explanation results on MNIST dataset [10], where we can compare to human judgment about the feature importance. We trained a convolutional network (Conv2D(32) → Conv2D(64) → Dense(128) → Dense(10)) achieving 97.32% accuracy. We parcellate the image into ROIs using *slic* [11] to mimic the setting



a) Original image b) C-SVE c) H-SVE d) Single ROI exam

Fig. 2: The predictive power for identifying (a) the digit 8 by b) C-SVE, c) H-SVE, and d) single ROI explanation. The prediction difference after corrupting the ROIs which contribute 90% in total are denoted on the left corner.

of detecting saliency brain ROI for identifying ASD. Denoting the distance between the center of ROI i and ROI j as d_{ij}, we define the connection between ROI i and j as $a_{ij} = exp(-d_{ij}/2)$. Here we use $th = \frac{\sum_i \sum_j a_{ij}}{|\mathcal{E}|}$. The results are shown in Fig. 2, where we uniformly divided each ROI's importance score by the number of pixels in the ROI to mitigate dominance by large ROIs and divided by $\max_{i \in \mathcal{N}}(\Phi_i)$ for visualization. The interpretation results matched our human perception that the "*x cross*" shape in the center is important for recognizing digit 8. Compared with single ROI testing, our proposed methods assigned smoother and more widely distributed importance scores to more pixels. To examine the effect of important ROIs on prediction, we corrupted pixels whose importance power added up to 90% of the positive importance scores. We then compared the difference between the original prediction probability of digit 8 and the new prediction probability using the corrupted image. C-SVE and H-SVE could better fool the classifier, which decreased the prediction probability by 0.8939 and 0.9089 respectively, compared to only a 0.2043 decrease for the single ROI method. Some ROIs may not contribute to classification on their own but influence the results when combined with other regions. In the single ROI method, these ROIs will be assigned 0 importance score. However, by our proposed SVE method these ROIs can be discovered.

4.2 ASD Task-fMRI Dataset and Underlying Graph Structure

We tested our methods on a group of 82 children with ASD and 48 age and IQ-matched healthy controls used for training the classifiers. Each subject underwent a biological motion perception task [3] fMRI scan (BOLD, TR = 2000ms, TE = 25ms, flip angle = 60°, voxel size $3.44 \times 3.44 \times 4mm^3$) acquired on a Siemens MAGNETOM Trio TIM 3T scanner. We randomly split 80% of the data for training, 10% for validation of model parameters, and 10% for testing.

The Automated Anatomical Labeling (AAL) atlas [12] was used to parcellate the brain into 116 regions. For each subject, we computed the 116×116 adjacency matrix using Pearson correlation. We averaged the adjacency matrix over the patient subjects in the training data and binarized the edges based on whether its weight is larger than average weight (assigning 1) or not (assigning 0). For H-SVE method, we obtained the non-overlapping community clustering for each subject by greedy modularity method [13], which resulted in 10 communities.

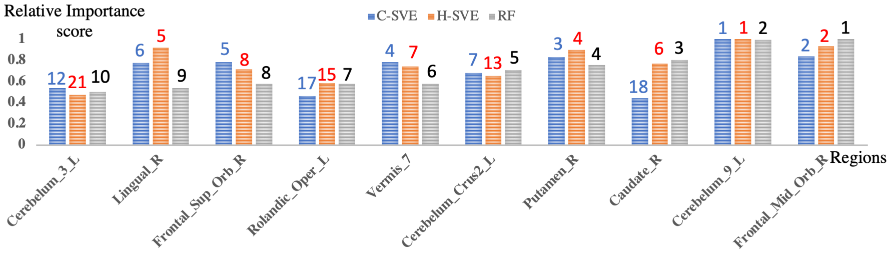


Fig. 3: The relative importance scores of the top 10 important ROIs assigned by Random Forest and their corresponding importance scores in C-SVE and H-SVE. The importance rank of each ROI is denoted on the bar.

4.3 Comparison with Random Forest-based Feature Importance

As an additional "reality check" for our method, we apply a Random Forest (RF) strategy (1000 trees) to the same dataset (71.4% accuracy on testing set) and compare the results, using the RF-based feature importance (mean Gini impurity decrease) as a form of standard method for comparison. Instead of inputting the entire fMRI image, we input the node-weighted modularity, which is defined by $\mathcal{M}_i = \sum_{j \neq i} a_{ij}$ where a_{ij} is the partial correlation coefficient between ROI i and j. Therefore the inputs are 116×1 vectors. Based on axiom 4, we can treat each subject as a game and each ROI as a player, and then do group-based analysis by adding $\Phi(r)$ over the subjects to investigate ROI r's importance. For a fair comparison, like in RF, we used all of the training dataset. The interpretation results are shown in Fig. 3. Seven of the top 10 important ROIs discovered by C-SVE and H-SVE overlapped with RF interpretation.

4.4 Explaining The ASD Brain Biomarkers Used In Deep Convolutional Neural Network Classifier

Here we chose the deep neural network 2CC3D (Fig. 4) described in [14] using each voxel's mean and standard deviation as two channel input. We start with preprocessed 3D fMRI volumes downsampled to $32 \times 32 \times 32$. We defined the original fMRI sequence as $\boldsymbol{X}(x, y, z, t)$, the mean-channel sequence as $\tilde{\boldsymbol{X}}(x, y, z, t)$ and

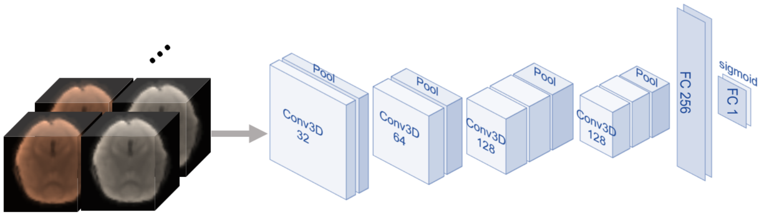


Fig. 4: 2CC3D network architecture

Table 1: Prediction Decrease After Corrupting Important ROIs for the DNN

	C-SVE	H-SVE	Single Region
$\Delta prob$	0.720 (0.221)	0.693 (0.144)	0.335 (0.060)
Δacc	0.714	0. 714	0.428

($\Delta prob$ = decrease in test prediction probability, Δacc = decrease in test accuracy)

the standard deviation-channel as $\hat{\boldsymbol{X}}(x,y,z,t)$. For any x,y,z in $\{0,1,\cdots,31\}$,

$$\tilde{\boldsymbol{X}}(x,y,z,t) = \frac{\sum_{\tau=t+1-w}^{t} \boldsymbol{X}(x,y,z,\tau)}{w} \tag{9}$$

$$\hat{\boldsymbol{X}}(x,y,z,t) = \sqrt{\frac{\sum_{\tau=t+1-w}^{t}[\boldsymbol{X}(x,y,z,\tau) - \tilde{\boldsymbol{X}}(x,y,z,t)]^2}{w-1}}, \tag{10}$$

where w is the temporal sliding window size. It achieved 85.7% classification accuracy when $w = 3$ on the task-fMRI dataset. Running on a workstation with a Nvidia 1080 Ti GPU, testing all 7 ASD subjects in the testing dataset took $21k$ s and $26k$ s for C-SVE and H-SVE, respectively, using 1000 samples for MC approximation, which converged to the stable ranks. As in the MNIST experiment, we divided $\Phi(r)$ by the number of voxels in ROI r, avoiding domination by large ROIs.

The contribution/prediction power of the regions (relative to the most important one) averaged over testing subjects are illustrated in Fig. 5 and listed in Fig. 6. There are 19 overlapping ROIs out of the top 20 important ROIs found by C-SVE and H-SVE, although the orders were different. The *Spearman rank-order correlation coefficient* [15] of the importance score ranks of all the ROIs explained by both methods was 0.58. These detected regions were consistent with the previous findings in the literature [2,3]. Also, we used Neurosynth [16] to decode the functional keywords associated with the overlapping biomarkers found by C-SVE and H-SVE (Fig. 7). These top regions are positively related to self-referential/perspective-taking concepts (higher level social communication) and

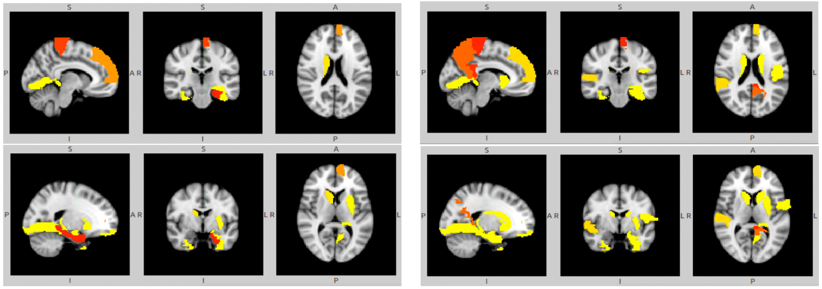

a) C-SVE
b) H-SVE

Fig. 5: Top 20 predictive biomarkers detected by a) C-SVE and b) H-SVE for the deep learning classifier. More yellow ROIs signify higher importance.

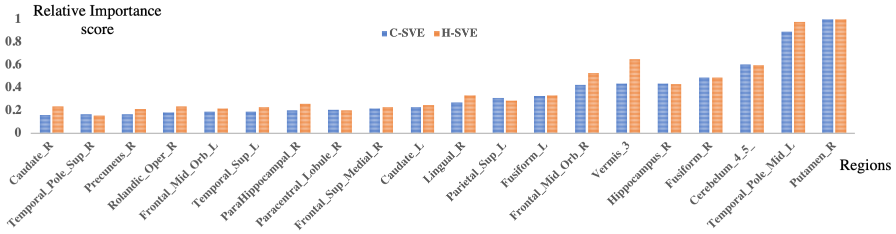


Fig. 6: The relative importance scores of the top 20 ROIs assigned by C-SVE and their corresponding importance scores in H-SVE for the deep learning model.

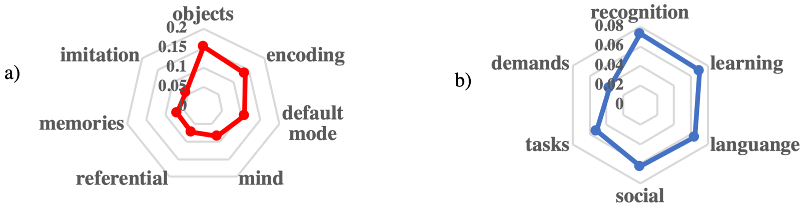


Fig. 7: a) The top *positive* correlations and b) the top *negative* correlations between deep learning model biomarkers and functional keywords.

negatively related to more basic social and language concepts (lower level skills). Using the manner described in Eq. (1), we corrupted the important ROIs (50% of the positive importance scores summing up in order) determined by C-SVE, H-SVE, and single region testing separately and calculated the average decrease in probability $\Delta prob$ (showing mean and standard deviation) and accuracy Δacc for the subjects in the testing set. The results are listed in Table 1.

Notice that the top 10 biomarkers we discovered using SVE in the RF model were different from the ones found in the 2CC3D model. Possible reasons are: 1) the inputs are different. 2CC3D used activation whereas RF used connectivity and 2CC3D used ASD subjects in testing set whereas RF used all the training set; 2) the prediction accuracy of RF model is much lower than 2CC3D; and 3) our proposed methods performed as a model interpreter rather than data interpreter, which may have different sensitivity response to the different models.

5 Conclusion And Future Work

Considering the interaction of features, we proposed two approaches (C-SVE and H-SVE) to analyze feature importance based on SVE, using the underlying graph structure of the data to simplify the calculation of Shapley value. C-SVE only considers the centralized interaction, while H-SVE uses a hierarchical approach to first cluster the feature communities, then calculate the Shapley value in each community. When a feature's neighborhood/community still contains a large number of features, we apply MC integration method for further approximation. Experiments on the MNIST dataset showed our proposed methods can capture more interpretable features. Comparing the results with Random Forest feature interpretation on the ASD task-fMRI dataset, we further validated the accuracy and feasibility of the proposed methods. When applying both methods on a deep learning model, we discovered similar possible brain biomarkers, which matched the findings in the literature and had meaningful neurological interpretation. The pipeline can be generalized to other feature importance analysis problems, where the underlying graph structure of features is available.

Our future work includes testing the methods on different atlases, graph building methods, and community clustering methods, etc. In addition, the interaction score is embedded in the proposed algorithms. It can be disentangled to understand the interaction between the features.

A Appendix: Proof of Theorem 2

For any subset $A \subset \mathcal{N}$, we use the short notation $U_r(A) := A \cap \mathcal{N}(r)$ and $V_r(A) := A \cap (\mathcal{N} \setminus \mathcal{N}(r))$, noting that $A = U_r(A) \cup V_r(A)$. Denoting $\Delta_r^{\boldsymbol{X}}(U, A) = (v_{\boldsymbol{X}}(U \cup \{r\}) - v_{\boldsymbol{X}}(U)) - (v_{\boldsymbol{X}}(A \cup \{r\}) - v_{\boldsymbol{X}}(A))$, then we have

$$\Delta_r^{\boldsymbol{X}}(U, A) = \log \frac{p(Y|X_{\mathcal{N}\setminus U})}{p(Y|X_{\mathcal{N}\setminus(U\cup\{r\})})} - \log \frac{p(Y|X_{\mathcal{N}\setminus(U\cup V)})}{p(Y|X_{\mathcal{N}\setminus(U\cup V\cup\{r\})})} \tag{11}$$

Abbreviating $V_r(A)$ as V, let $W = \mathcal{N} \setminus (\mathcal{N}(r) \cup V)$, $Z = \mathcal{N}(r) \setminus (\{r\} \cup U)$. Then

$$\Delta_r^{\boldsymbol{X}}(U, A) = \log \frac{p(Y|X_{W\cup V\cup Z\cup\{r\}})p(Y|X_{W\cup Z})}{p(Y|X_{W\cup V\cup Z})p(Y|X_{W\cup Z\cup\{r\}})} \tag{12}$$

We have $p(X_V|X_{W\cup Z\cup\{r\}}) = p(X_V|X_{W\cup Z})$, since $X_r \perp X_V|X_Z$. Then $(\star) = \frac{p(X_{W\cup V\cup Z\cup\{r\}})p(X_{W\cup Z})}{p(X_{W\cup V\cup Z})p(X_{W\cup Z\cup\{r\}})} = \frac{p(X_{W\cup Z\cup\{r\}})p(X_V|X_{W\cup Z\cup\{r\}})p(X_{W\cup Z})}{p(X_{W\cup Z})p(X_V|X_{W\cup Z})p(X_{W\cup Z\cup\{r\}})} = 1$. Thus, we can multiply the quotient in Eq. (12) by $(\star)$:

$$\Delta_r^{\boldsymbol{X}}(U, A) = \log \frac{p(Y|X_{W\cup V\cup Z\cup\{r\}})p(Y|X_{W\cup Z})}{p(Y|X_{W\cup V\cup Z})p(Y|X_{W\cup Z\cup\{r\}})} \frac{p(X_{W\cup V\cup Z\cup\{r\}})p(X_{W\cup Z})}{p(X_{W\cup V\cup Z})p(X_{W\cup Z\cup\{r\}})} \tag{13}$$

$$= \log \frac{p(X_{W\cup V\cup\{r\}}|Y, X_Z)p(Y, X_Z)p(Y, X_Z)p(X_W|Y, X_Z)}{p(Y, X_Z)p(X_{W\cup V}|Y, X_Z)p(Y, X_Z)p(X_{W\cup\{r\}}|Y, X_Z)}. \tag{14}$$

We have $p(X_{W\cup V\cup\{r\}}|Y, X_Z) = p(X_{W\cup V}|Y, X_Z)p(X_r|Y, X_Z)$, since $(W \cup V) \perp \{r\}|Y, X_Z$. So

$$\Delta_r^{\boldsymbol{X}}(U, A) = \log \frac{p(X_{W\cup V}|Y, X_Z)p(X_r|Y, X_Z)p(X_W|Y, X_Z)}{p(X_{W\cup V}|Y, X_Z)p(X_{W\cup\{r\}}|Y, X_Z)}. \tag{15}$$

Since $W \perp \{r\}|Y, X_Z$, we have $p(X_{W\cup\{r\}}|Y, X_Z) = p(X_W|Y, X_Z)p(X_r|Y, X_Z)$. Hence $\Delta_r^{\boldsymbol{X}}(U, A) = \log 1 = 0$. Rewrite equations (4) and (2) as

$$\hat{\Phi}_r^C(v_{\boldsymbol{X}}) = \frac{1}{|\mathcal{N}(r)|} \sum_{U\subseteq\mathcal{N}(r)\setminus\{r\}} \binom{|\mathcal{N}(r)| - 1}{|U|}^{-1} (v_{\boldsymbol{X}}(U \cup \{r\}) - v_{\boldsymbol{X}}(U)) \tag{16}$$

$$\Phi_r(v_{\boldsymbol{X}}) = \frac{1}{|\mathcal{N}|} \sum_{U\subseteq\mathcal{N}(r)\setminus\{r\}} \sum_{A\subseteq\mathcal{N}, U_r(A)=U} \binom{|\mathcal{N}|-1}{|A|}^{-1} (v_{\boldsymbol{X}}(A\cup\{r\}) - v_{\boldsymbol{X}}(A)), \tag{17}$$

then the expected error between $\hat{\Phi}^C_r(v_{\boldsymbol{X}})$ and $\Phi_r(v_{\boldsymbol{X}})$ is

$$\mathbb{E}[|\hat{\Phi}^C_r(v_{\boldsymbol{X}}) - \Phi_r(v_{\boldsymbol{X}})|] \leq \frac{1}{|\mathcal{N}|} \sum_{U\subseteq\mathcal{N}(r)\setminus\{r\}} \sum_{A\subseteq\mathcal{N}, U_r(A)=U} \binom{|\mathcal{N}|-1}{|A|}^{-1} \mathbb{E}[|\Delta^{\boldsymbol{X}}_r(U, A)|] \tag{18}$$

where we use $\sum_{A\subseteq\mathcal{N}, U_r(A)=U} \binom{|\mathcal{N}|-1}{|A|}^{-1} = \frac{|\mathcal{N}|}{|\mathcal{N}(r)|} \binom{|\mathcal{N}(r)|-1}{|U|-1}^{-1}$. Therefore we have $\mathbb{E}[|\hat{\Phi}^C_r(v_{\boldsymbol{X}}) - \Phi_r(v_{\boldsymbol{X}})|] = 0$.

References

1. A. A. Goldani, S. R. Downs, F. Widjaja, B. Lawton, and R. L. Hendren, "Biomarkers in autism," *Frontiers in psychiatry*, vol. 5, 2014.
2. X. Li, N. C. Dvornek, J. Zhuang, P. Ventola, and J. S. Duncan, "Brain biomarker interpretation in asd using deep learning and fmri," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 206–214, Springer, 2018.
3. M. D. Kaiser, C. M. Hudac, S. Shultz, S. M. Lee, C. Cheung, A. M. Berken, B. Deen, N. B. Pitskel, D. R. Sugrue, A. C. Voos, *et al.*, "Neural signatures of autism," *Proceedings of the National Academy of Sciences*, vol. 107, no. 49, pp. 21223–21228, 2010.
4. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, pp. 4765–4774, 2017.
5. J. Chen, L. Song, M. J. Wainwright, and M. I. Jordan, "L-shapley and c-shapley: Efficient model interpretation for structured data," *arXiv preprint arXiv:1808.02610*, 2018.
6. I. Kononenko *et al.*, "An efficient explanation of individual classifications using game theory," *Journal of Machine Learning Research*, vol. 11, no. Jan, pp. 1–18, 2010.
7. L. S. Shapley, "A value for n-person games," *Contributions to the Theory of Games*, vol. 2, no. 28, pp. 307–317, 1953.
8. L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," *arXiv preprint arXiv:1702.04595*, 2017.
9. A. Clauset, M. E. Newman, and C. Moore, "Finding community structure in very large networks," *Physical review E*, vol. 70, no. 6, p. 066111, 2004.
10. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
11. R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, *et al.*, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

12. N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, “Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain,” *Neuroimage*, 2002.
13. M. Newman, *Networks*. Oxford university press, 2018.
14. X. Li, N. C. Dvornek, X. Papademetris, J. Zhuang, L. H. Staib, P. Ventola, and J. S. Duncan, “2-channel convolutional 3d deep neural network (2cc3d) for fmri analysis: Asd classification and feature learning,” in *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*, pp. 1252–1255, IEEE, 2018.
15. R. C. Young, J. T. Biggs, V. E. Ziegler, and D. A. Meyer, “A rating scale for mania: reliability, validity and sensitivity,” *The British journal of psychiatry*, vol. 133, no. 5, pp. 429–435, 1978.
16. T. Yarkoni, R. A. Poldrack, T. E. Nichols, D. C. Van Essen, and T. D. Wager, “Large-scale automated synthesis of human functional neuroimaging data,” *Nature methods*, vol. 8, no. 8, p. 665, 2011.