Fall 12-14-2022

Decentralized Harmonization Algorithm and Application to Functional Network Connectivity

Biozid Bostami

Decentralized Harmonization Algorithm and Application to Functional Network Connectivity

by

Biozid Bostami

Under the Direction of Vince Calhoun, PhD

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2022

ABSTRACT

Neuroimage data collected from multiple research institutions may incur additional source dependency, affecting the overall statistical power and leading to erroneous conclusions. This problem can be mitigated with data harmonization approaches. While open neuroimaging datasets are becoming more common, a substantial amount of data can still not be shared for various reasons. In addition, current approaches require moving all the data to a central location, which requires additional resources and creates redundant copies of the same datasets. To address these issues, we propose a decentralized harmonization approach called "Decentralized ComBat" that performs remote operations on the datasets separately without sharing individual subject data, ensuring a certain level of privacy and reducing regulatory hurdles. The study was conducted on harmonizing functional connectivity. Results showed similar performance as the centralized ComBat algorithm in a decentralized environment.

INDEX WORDS: Harmonization, Neuroimaging, Machine Learning

Decentralized Harmonization Algorithm and Application to Functional Network Connectivity

by

Biozid Bostami

Committee Chair: Vince Calhoun

Committee: Jingyu Liu

Murray Patterson

Electronic Version Approved:

Office of Graduate Services

College of Arts and Sciences

Georgia State University

December 2022

DEDICATION

This thesis is dedicated to all the teachers who appeared at different of my life and guided me to keep learning and growing. I also honor everyone who has dedicated their lives to the causes of research and making this world a better place. I also want to dedicate my work to my parents and siblings, who have been my most vital support throughout my life. I also want to mention all my friends who have constantly given me the courage to pursue and persist for the best.

ACKNOWLEDGEMENTS

I want to express my gratitude to Dr. Vince Calhoun for mentoring me through my graduate journey and introducing me to academic research experience in neuroscience. His patient and dynamic approach allowed me to develop and channel my abilities toward purposeful research relying on conscious choices and common sense-based thinking. I thank him for supporting me through the difficult phases of this research. I also want to thank Dr. Victor Vergara, who helped me through the project and learning process. Without these kindhearted people, I would not be able to accomplish this.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS V

LIST OF FIGURES VIII

1 INTRODUCTION 1

1.1 Motivation 1

1.2 Contributions 3

2 METHODS 4

2.1 ComBat 4

3 DECENTRALIZED COMBAT MODEL OVERVIEW: 5

4 DATA COLLECTION AND PRE-PROCESSING 8

4.1 Dataset 8

4.1.1 New Mexico Cohort Imaging Protocol 9

4.1.2 Netherland (European) Cohort Imaging Protocol 9

4.1.3 Data Pre-Processing 10

4.2 Dataset 2 10

5 EXPERIMENTAL SETUP AND OBSERVATIONS 11

5.1 Experimental Setup 1 and Observations 11

5.2 Experimental Setup 2 and Observations 13

5.3 Experimental Setup 3 and Observations 15

6 DISCUSSION 17

7 LIMITATIONS AND FUTURE DIRECTIONS 20

8 CONCLUSION 21

REFERENCES.. 21

LIST OF FIGURES

Figure 1 Gives the overall picture of the decentralized ComBat algorithm and intra-communication between nodes. 8

Figure 2 Heatmap of t-values site-difference (NM-EU) before (left) and after harmonization (right). 12

Figure 3 Heatmap of t-values group difference (NM-EU) before (left) and after harmonization (right). 13

Figure 4 Decentralized ComBat with different distributions as the multiplicative parameter; Gaussian distribution (skewness: 0.26 and kurtosis: 3.3) (left) and Sub-Gaussian distribution (skewness: 0.12 and kurtosis: 2.2) (right) for additive parameter. 14

Figure 5 Decentralized ComBat with different distributions as the multiplicative parameter; Skewed-left distribution(skewness: -0.58 and kurtosis: 3.48) (left) and right skewed distribution (skewness: 0.34 and kurtosis: 2.34) (right) for additive parameter 15

Figure 6 AUC score comparison of different classifier for unharmonized and harmonized datasets considering all features 17

Figure 7 AUC score comparison of different classifier for unharmonized and harmonized datasets considering feature selection. 17

1 INTRODUCTION

1.1 Motivation

The significance of neuroscience studies has reached a global scale with an increasing number of large-scale projects related to impactful topics such as brain disease, brain development, brain aging, and brain-computer interfacing [1,2,3]. The true potential of these large projects depends on the data at one's disposal, which urges global collaboration, knowledge, and data sharing. These collaborative approaches include aggregating data collection to a central repository or data sharing based on data usage agreements (DUA) [4,5]. Such an approach has several limitations to consider. The first concern is the policy and proprietary restrictions, or data de-identification issues may be raised. Such concerns are time-consuming and take months to resolve.

Moreover, the processing of DUA may take months for approval. Although many open datasets may be quickly accessible without DUA, that brings us to our next concern. Another significant concern is the volume of the data collected from multiple sites. When we merge multiple large neuroimage datasets in a single location, it will consume much space. Additionally, computational resources become costly when the volume of data grows. Also, sharing the data only creates redundant copies around the world. Thus, it is not always an optimal approach considering the constraints on available resources. While open neuroimage datasets are becoming more common, some data cannot be transferred or shared directly due to confidentiality or regulatory constraints. These issues led to a paradigm shift towards decentralized data-sharing [6,7], particularly with widespread efforts in the neuroimaging community to maximize study power through multi-site investigation, data sharing, and team science.

With the availability of neuroimage data at multiple sites worldwide, an important goal is to jointly analyze geographically dispersed data to increase statistical power and test against the common biological hypothesis. There is an issue with combining the multi-site neuroimage data because each data at a different location introduces additional non-biological variability. These variabilities are closely related to image acquisition protocol and scanner parameters categorized as 'site effects' [8]. These site effects can reduce statistical power or lead to erroneous conclusions. Harmonization techniques aim to combine datasets generated from different sites, e.g., hospitals, research facilities, or laboratories, reducing the site effects in the combined dataset [9].

One popular harmonization technique is known as ComBat [10]. The ComBat technique was first introduced in genomics to reduce batch effects and non-biological variability due to pooling batches of sample genes from various laboratories. Later, it was applied to diffusion tensor imaging (DTI) [9], cortical thickness data [11], functional connectivity measures [12], and positron emission tomography (PET) imaging [13]. However, the current ComBat model does not address data access problems, including geographical and confidentiality issues, which motivated us to develop a decentralized model that works in a distributed environment. This manuscript presents a decentralized harmonization model called 'Decentralized ComBat (DC-ComBat).'

Including decentralized regression [14], decentralized temporal independent component analysis [15], decentralized independent vector analysis [16], decentralized neural networks [17], decentralized data ICA [18], decentralized PCA [19] and many more. Some of these algorithms can be used jointly with our decentralized harmonization approach in the COINSTAC for creating different pipelines. Based on the benefits, we found this framework suitable for our decentralized approach.

1.2 Contributions

Our team has been working on a web-based framework for several years to analyze data stored in multiple locations without pooling, named Collaborative Informatics and Neuroimaging Suite Toolkit for Anonymous Computation (COINSTAC) [14]. This framework also preserves the privacy of the data as there is no data pooling involved, and all the communication between the sites is encrypted. COINSTAC uses a message-passing infrastructure to implement decentralized algorithms to work with geographically scattered datasets. We can develop a decentralized algorithm with this framework that returns similar results on collected datasets. This framework preserves dataset privacy by not creating additional copies. Also, this framework can be scaled easily when the number of sites or datasets increases. There are several decentralized algorithms already implemented using COINSTAC. Some of the decentralized computations proposed earlier

2 METHODS

2.1 ComBat

ComBat can be described as follows if the data is collected from $\boldsymbol{k}$ different sites where each site has $\boldsymbol{n_i}$ scans where $\boldsymbol{i}$ = 1, 2, …, $\boldsymbol{k}$. Each harmonized feature $\boldsymbol{y}$ indexed by $\boldsymbol{v}$ of scan $\boldsymbol{j}$ at site $\boldsymbol{i}$, the value $\boldsymbol{y_{i,j,v}}$ can be defined as:

$$\boldsymbol{y_{i,j,v}} = \boldsymbol{\alpha_v} + \boldsymbol{X_{i,j}\beta_v} + \boldsymbol{\gamma_{i,v}} + \boldsymbol{\delta_{i,v}\varepsilon_{i,j,v}} \qquad (1)$$

In the above equation $\boldsymbol{\alpha_v}$ represents the overall mean value at feature $\boldsymbol{v}$. $\boldsymbol{X}$ represents the biological variants, $\boldsymbol{\beta_v}$ represents the regression coefficient for **X** at feature $\boldsymbol{v}$. The error term **ε** is assumed to follow a Gaussian distribution $\mathbf{N(0,\boldsymbol{\sigma}^2)}$. In equation(1) $\boldsymbol{\delta_{i,v}}$ and $\boldsymbol{\gamma_{i,v}}$ represents the multiplicative and additive parameters correcting for site effects at site $\boldsymbol{i}$ for feature $\boldsymbol{v}$. The model aims to reduce the unwanted variance using the Empirical Bayes approach. The final distribution model can be achieved by:

$$\boldsymbol{y_{ijv}^{comBat}} = \frac{\boldsymbol{y_{ijv} - \hat{\alpha}_v - X_{ij}\hat{\beta}_v - \hat{\gamma}_{iv}}}{\boldsymbol{\hat{\delta}_{iv}}} + \boldsymbol{\hat{\alpha}_v} + \boldsymbol{X_{ij}\hat{\beta}_v} \qquad (2)$$

The model can be divided into three parts. The first part is the standardization of data. After standardization, every data will have similar overall mean and variance. The following equation calculates the standardization data:

$$\boldsymbol{Z_{i,j,v}} = \frac{\boldsymbol{y_{ijv} - \hat{\alpha}_v - X_{ij}\hat{\beta}_v}}{\boldsymbol{\hat{\sigma}_v}} \qquad \mathbf{(3)}$$

The second part is the estimation of batch effect using parametric empirical priors. The ComBat assumes that the standardized data $\boldsymbol{Z_{i,j,v}}$ follows the standard distribution form, $\boldsymbol{Z_{i,j,v}} \sim \mathbf{N}\left(\boldsymbol{\gamma_{i,v}}, \boldsymbol{\delta_{i,v}^2}\right)$. It is also mentioned that parametric forms of the prior distributions on the batch effect parameters, $\boldsymbol{\gamma_{i,v}}$, $\boldsymbol{\delta^2_{i,v}}$ follows a normal distribution and Inverse gamma distribution, respectively. Defined by:

$$\boldsymbol{\gamma_{i,v}} \sim \mathbf{N}\left(\boldsymbol{Y_i}, \boldsymbol{\tau_i^2}\right) \quad \text{and} \quad \boldsymbol{\delta^2_{i,v}} \sim \textbf{Inverse Gamma}(\boldsymbol{\lambda_i}, \boldsymbol{\theta_i}) \qquad \textbf{(4)}$$

The hyperparameters $\boldsymbol{\gamma_i}, \boldsymbol{\tau_i^2}, \boldsymbol{\lambda_i}, \boldsymbol{\theta_i}$ are estimated empirically from the standardized data. Details of the derivation of the estimators are explained in the supplementary material of the original ComBat paper [8]. Based on the Empirical Bayes estimators $\gamma_{i,v}, \delta^2_{i,v}$ can be defined by the posteriors means as followings:

$$\boldsymbol{\gamma^*_{i,v}} = \frac{\boldsymbol{n_i \tau_i^2 \widehat{\gamma_{\imath,v}} + \delta^{2*}_{i,v} \overline{\gamma_\imath}}}{\boldsymbol{n_i \tau_i^2 + \delta^{2*}_{i,v}}} \qquad \text{and} \qquad \boldsymbol{\delta^{2*}_{i,v}} = \frac{\boldsymbol{\overline{\theta_\imath} + \frac{1}{2} \sum_j \left(Z_{ijv} - \gamma^*_{i,v}\right)^2}}{\boldsymbol{n_i \tau_i^2 + \delta^{2*}_{i,v}}} \qquad \textbf{(5)}$$

Finally, data is adjusted based on the estimated site parameters $\boldsymbol{\gamma^*_{i,v}}$ $\boldsymbol{and}$ $\boldsymbol{\delta^{2*}_{i,v}}$.

The described ComBat model does not address working in a decentralized environment. We proposed a decentralized model that can operate on separate datasets and produce identical results to the original model. We implemented the decentralized ComBat (DC-ComBat) using a platform COINSTAC. The architecture of DC-ComBat- is discussed in the following section.

3 DECENTRALIZED COMBAT MODEL OVERVIEW:

In our decentralized environment, we have two types of nodes: The first type is the aggregator node, also known as the remote node, which does not hold any data and acts as a storage of intermediate results and performs simple operations such as aggregation. The second node type is the local/regional node where datasets are located. These local nodes represent the participants who are willing to work collaboratively. With the help of COINSTAC, we created a network where the regional nodes can be connected to the remote node and perform different operations synchronously.

For harmonizing distributed datasets located at different locations, we first constructed a network prototype shown in Figure 1, where all the participating local nodes connect with the remote node. Then each participating local node shares the local number of samples with the remote node via the secured message-passing mechanism. All intermediate communication is encrypted and sent over TLS (Transport Layer Security) provided by COINSTAC [14]. Then the remote node calculates the total sample count across the participating nodes depicted in Figure 1(a). Then the remote node broadcasts the total sample count to all the participating local nodes Figure 1(b). After that, using the information about the total number of samples, we calculate the mean across the features at each site and the variance across the features at each site with respect to the total number of samples across the participating site nodes. Then each local node sends the calculated local mean and variance to the remote node. The remote node calculates the grand mean and grand variance by aggregating the regional nodes' values in Figure 1(c) and broadcasting the grand mean and grand variance to all local nodes in Figure 1(d). After receiving the grand mean and grand variance information from the remote node, each node performs data standardization on the dataset located at each node Figure 1(d). Following the data standardization, estimation of site effect using parametric empirical priors is done on each site. Moreover, each site can adjust and harmonize the local data concerning the other participating site nodes based on the estimated site parameters. The pseudo algorithm is given below:

Algorithm:

Step 1: Initialize the central node and site nodes.

Step 2: Collect the initial summary (number of samples) of the site nodes in the central node.

Step 3: Calculate the β coefficient for each site using the decentralized regression approach available in COINSTAC.

Step 4: For site node, i = 1, 2, 3 …. N do

1. calculate the local mean across the features using the local β coefficients
2. calculate the local variance across the feature using the local β coefficients
3. send the local mean and variance to the aggregator node.
4. End for loop.

Step 5: compute grand mean and grand variance and update each site node.

Step 6: For site node, i = 1, 2, 3 …. N, do

1. standardize the data w.r.t the grand mean and grand variance.
2. Estimate the site parameters $\gamma_{i,v}^{*}\ and\ \delta_{i,v}^{2*}$.
3. Adjust the data accordingly.
4. Save the adjusted data.
5. End for loop.

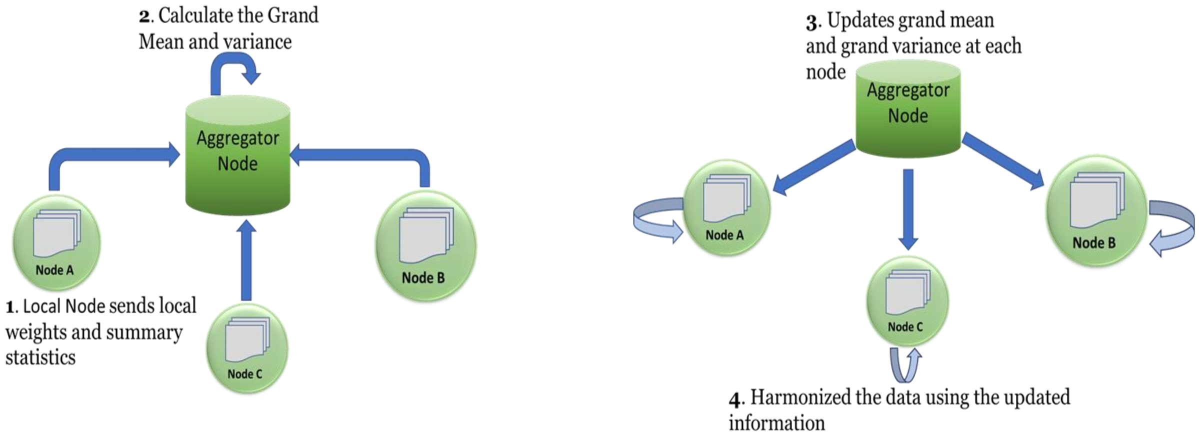
2. Calculate the Grand Mean and variance
Aggregator Node
Node A
Node B
Node C
1. Local Node sends local weights and summary statistics
3. Updates grand mean and grand variance at each node
Aggregator Node
Node A
Node B
Node C
4. Harmonized the data using the updated information

Figure 1 Gives the overall picture of the decentralized ComBat algorithm and intra-communication between nodes.

4 DATA COLLECTION AND PRE-PROCESSING

We used two sets of data for experimenting with our decentralized ComBat model. The first set consists of static FNC (functional network connectivity) data collected from two studies on mild traumatic brain injuries (mTBI) [20]. We wanted to observe our model's performance when applied to FNC data as from the previous study presented in [10], which showed that the ComBat model performs well in removing site effects from FNC datasets. The second set consists of simulated data generated using a connectivity template. The second dataset was used to measure the performance and scalability of our model when the number of sources increased. We tried to simulate a real-world situation where datasets located at different locations worldwide can be harmonized simultaneously. The following sections will describe how these two sets of datasets were collected and pre-processed.

4.1 Dataset

This dataset consists of data collected from two cohorts. The first cohort was collected from New Mexico (NM). All participants provided informed consent according to the Declaration of Helsinki and the institutional review board guidelines at the University of New Mexico. The second cohort was collected from the Netherlands Europe (EU). The local Medical Ethics Committee of the UMCG approved the data collection protocol, and every participant provided written informed consent. All procedures were conducted following the declaration of Helsinki. This data was also used in other studies related to dynamic functional connectivity [20] and brain modalities [21]. Data pre-processing and analysis were the same as described in the earlier research publication [22]; therefore, we present a brief outline of the whole process.

4.1.1 New Mexico Cohort Imaging Protocol

In the New Mexico cohort, the total number of participants was 96, among which 48 were mTBI patients and 48 were healthy control (HC). The subjects had a mean age of 27.3 ± 9.0 years. The scanner used in the New Mexico cohort was a 3 Tesla Siemens TIM Trio scanner. Every participant had gone through 5 min resting state-run. TR (Repetition Time) = 2000 ms; TE (Time of Echo) = 29 ms; flip angle = 75º; FOV (Field of View) = 240 mm; matrix size = 64 x 64. After removing the first five images due to the T1 equilibrium effect, the final 145 images were selected next step analysis.

4.1.2 Netherland (European) Cohort Imaging Protocol

In the case of the European cohort total of 74 participants were studied. There were 54 patients with mTBI and 20 Healthy controls among the participants. The mean age was 37, ranging from 19-64. The 3.0 T Philips Integra MRI scanner was used to collect the brain images for this group of participants. The duration was 10 min for the Netherlands cohort. TR (Repetition Time) = 2000 ms; TE (Time of Echo) = 20 ms; flip angle = 8º; FOV (Field of View) = 224 × 224 × 136.5 mm.

4.1.3 Data Pre-Processing

First, the fMRI data underwent Statistical Parametric Mapping (SPM) [23] and was transformed into Montreal Neurological Institute standard space. AFNI v17.1.03 software was used for de-spiking. The time courses were made orthogonal to 1) linear, quadratic, and cubic trends, 2) 6 realignment parameters, and 3) derivatives of realignment parameters. Data collected from the NM participants were used in the group independent component analysis (ICA) [24] using the GIFT software [25] to gather a set of functionally independent components. For Netherland cohort data, the group information guided ICA [26] (GIGICA) algorithm was used to match the 48 selected components. Finally, discarding the artifactual components, only 48 noise-free components were chosen as resting-state networks (RSNs) for further study.

4.2 Dataset 2

For this set, we generated data using computer simulation. The primary purpose of using a simulated dataset was to observe the scalability and performance of our model. Additionally, we used simulated data because the original ComBat model assumes that two site parameters: multiplicative and additive parameters drawn from the dataset, will follow inverse-gamma and gaussian distribution. However, in practice, such an assumption may not always hold. That is why we created a simulation where datasets may follow some other distribution, e.g., sub-gaussian distribution, super-Gaussian distribution, or a skewed distribution for additive parameters and Poisson, Rayleigh, or Weibull distribution for multiplicative parameters. To generate the datasets, we used an FNC (functional network connectivity) template based on an FNC matrix from a previous study [27] as the ground truth. We created various datasets by randomly adding site variance complying with the assumed normal and inverse gamma distributions. We fixed the Gaussian distribution parameters with the mean at 0.05 and the standard deviation at 0.3. For the inverse gamma distribution, we set the mean at 0.3 and the standard deviation at 0.5. We used this dataset to observe the performance of the DC-ComBat model.

5 EXPERIMENTAL SETUP AND OBSERVATIONS

We separated the experiments into three different parts. We used the actual datasets collected from two different sites in our first part. In the second experiment, we used simulated datasets. The first experiment aims to validate the results of our proposed decentralized ComBat method and the second experiment aims to measure the computational features of the decentralized ComBat method. The third experiment is to show how harmonization could improve the performance of machine algorithms. In the following three sections, we will describe each experiment and the observations separately.

5.1 Experimental Setup 1 and Observations

We keep two datasets collected from two research facilities into two local nodes for this experiment. We applied our model DC-ComBat to harmonize the datasets. We perform two assessments on the dataset to observe the harmonization performance. First, we compare the site differences before and after harmonization. So, we took the difference between the functional connectivity values of New Mexico(NM) and European(EU) sites, resulting in 1128 t-values. Instead of showing vectors, we converted them into a matrix where rows and columns represent each of the 48 ICA components, and the heatmap indicates the strength of the site difference. Figure 2 shows the site difference before and after harmonization. There were a high number of significant site differences before harmonization, observed in Figure 2(left). These indicate that site information added non-biological variance in the datasets, which is undesirable. After harmonization, we observed from Figure 2(right) that all the significant site differences were removed from the data. Removal of site differences indicates a high performance of DC-ComBat.

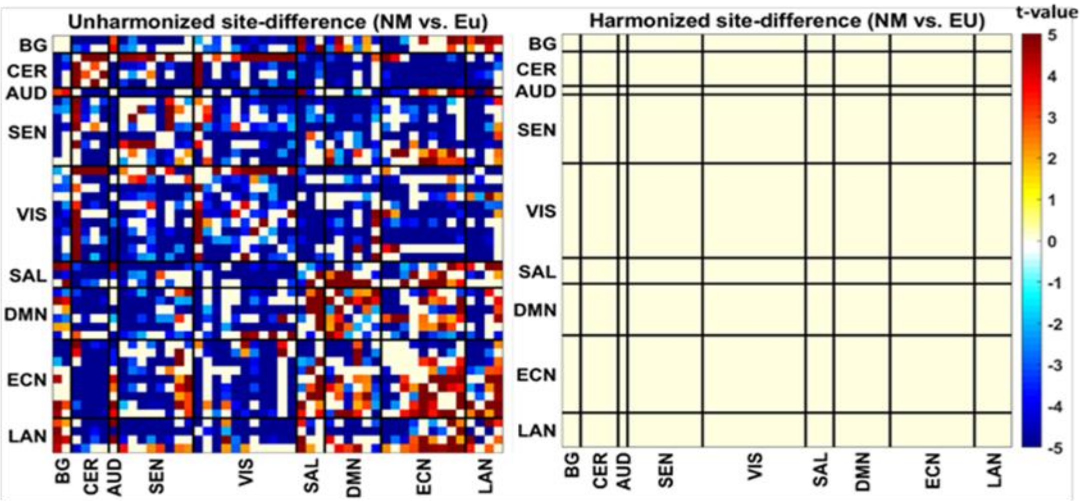

Figure 2 Heatmap of t-values site-difference (NM-EU) before (left) and after harmonization (right).

Later, we calculate the group difference (mTBI vs. HC) for the second assessment before and after harmonization. We first combined the datasets and calculated the group difference between participant groups (mTBI and Healthy Controls) before and after harmonization. Again, based on the t-values, we plot the heatmap shown in Figure 3. Before harmonization, there were 128 significant t-values ($p < 0.05$) shown in Figure 3(left); however, the number increased to 159 significant t-values when datasets were harmonized in Figure 3(right). After harmonization, higher connectivity was observed in the TBI group in general. We observed the increase in connectivity because, due to harmonization, site effects were posteriorly removed by DC-ComBat. Furthermore, by comparing the output of the proposed decentralized ComBat with centralized ComBat, it found that the maximum difference was $3.06699e^{-15}$. This slight difference in the output was within the order of magnitude of the machine precision error. We conclude that there was no practical difference between ComBat and DC-ComBat.

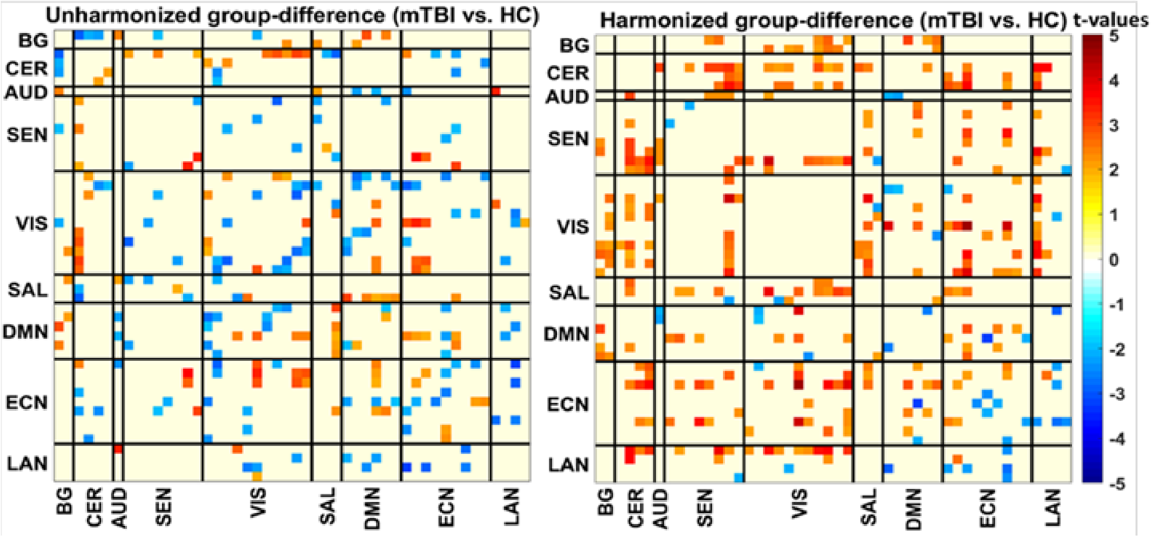

Figure 3 Heatmap of t-values group difference (NM-EU) before (left) and after harmonization (right).

5.2 Experimental Setup 2 and Observations

For the second experiment, we used simulation to generate data based on a functional connectivity template used as ground truth for further analysis [27]. We selected four probability distributions: Rayleigh, Weibull, Poisson, and inverse-gamma, to simulate the multiplicative parameter and added noise to the ground truth. Similarly, we selected Gaussian, Sub-gaussian, Right-skewed, and Left-skewed distributions for simulating additive site parameters and added noise to the ground truth. The selection of these probability distributions was random and without any prior knowledge. After adding the noise to the ground truth, we created several datasets. We created 250 datasets, each with 100 participants, random patients, and healthy controls. In the next step, we used COINSTAC-simulator to set the environment where each local node will contain a single dataset. Finally, we run our DC-ComBat algorithm to harmonize the datasets. We repeated the experiment by incrementing the number of sites and calculating the percentage of site effects removed with respect to the ground truth. The whole process was repeated four times by generating data with different distributions. We finally generated four plots in Figure 4 and Figure 5. The primary purpose of this experiment was to evaluate the consistency of our model when the number of sites increases and randomness is introduced. From Figure 4 and Figure 5, we observed that our model performance was not affected when the number of sites was more than 50. We did not observe any performance issues even when the number of sites increased, indicating that our proposed model is scalable and robust. The second purpose of using simulated data was to observe the performance of DC-ComBat when exposed to different site parameters drawn from different probability distributions. From Figure 4 and Figure 5, we observed that the skewness and kurtosis of the additive parameter affect the performance of the harmonization process. In Figure 5, we observed that when skewness and kurtosis increased, the algorithm could remove up to a maximum of 70% compared to Figure 4, where skewness and kurtosis were lower, and accuracy maximum accuracy was only 54%. We also observed from our experiment that the performance of DC-ComBat degrades for a certain distribution choice for multiplicative parameters. In Figure 4 and Figure 5, we saw that for the Poisson distribution, performance is poor compared to other distributions.

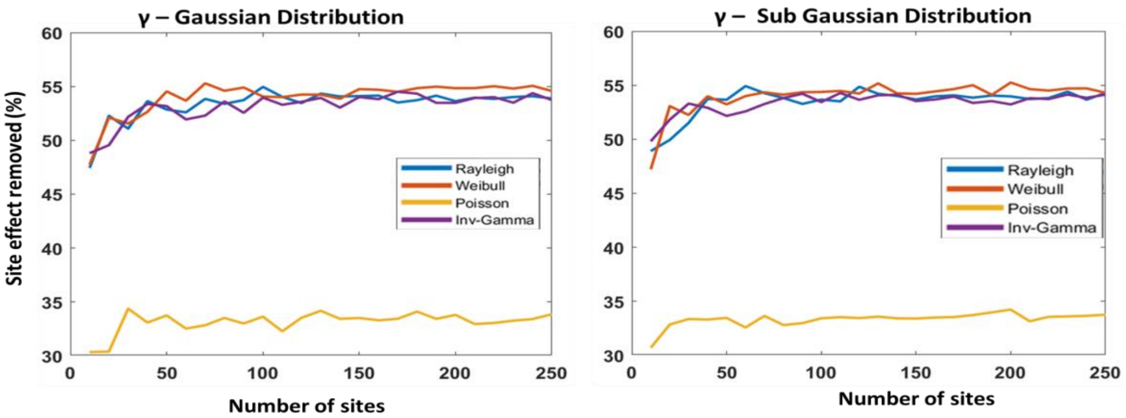

Figure 4 Decentralized ComBat with different distributions as the multiplicative parameter; Gaussian distribution (skewness: 0.26 and kurtosis: 3.3) (left) and Sub-Gaussian distribution (skewness: 0.12 and kurtosis: 2.2) (right) for additive parameter.

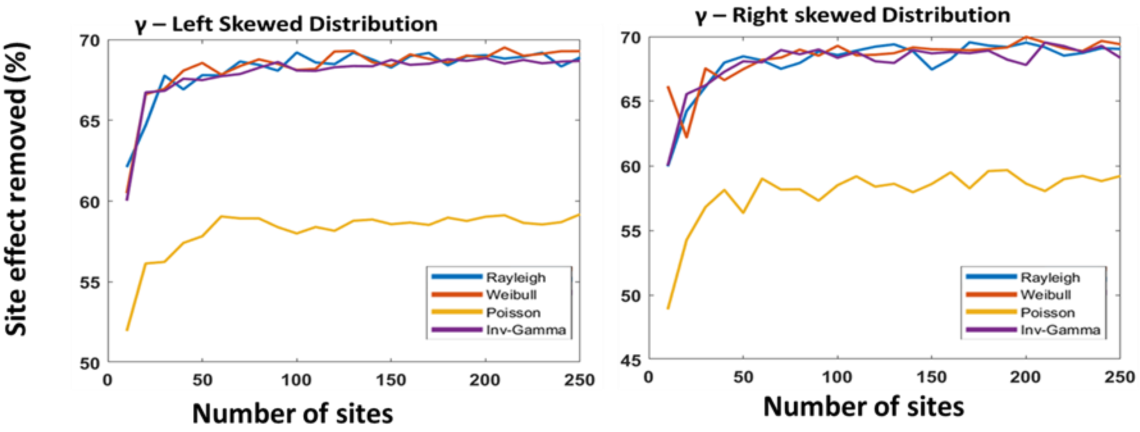

Figure 5 Decentralized ComBat with different distributions as the multiplicative parameter; Skewed-left distribution(skewness: -0.58 and kurtosis: 3.48) (left) and right skewed distribution (skewness: 0.34 and kurtosis: 2.34) (right) for additive parameter

5.3 Experimental Setup 3 and Observations

In this experiment, fMRI data were collected from two site sources. Scan data were preprocessed to obtain rsFNC values. Next, form a large dataset with sites' rsFNC data. Then, harmonized the combined dataset using the ComBat algorithm. To minimize any potential confounding influence of age and gender, linear regression was used to regress age and sex from rsFNC data. These residuals were further used as rsFNC data for the machine learning classifiers. The dataset was prepared for machine learning classifiers by splitting it into training and testing datasets (80:20). The training dataset was used to train the classifiers and the test data to evaluate their performances.

Next, for each classifier, a tuned model was built by performing grid-search 10-fold cross-validation, providing a set of hyperparameters and the training set as the input data. The model with the best area under the curve (AUC) average test score was selected as the classifier's tuned model. Feature selection is achieved by extracting the random forest feature importance values. The lower-dimensional features, referred to as selected features, are obtained by keeping the features with non-zero discriminative power. Then, the classifiers' tuned models were trained considering two input data cases: higher-dimensional data (all features) and lower-dimensional data (selected features). Using the AUC metric, the tuned models' performance was evaluated with the test data. The model with the best average AUC score was selected as the classifier's tuned model. Finally, tested the tuned models with the test dataset and reported the AUC scores. Also performed another experiment without applying the feature selection step in parallel and collected the results.

Similarly, repeated the analysis with unharmonized data and collected the results. Finally, plot the AUC scores of different classifiers for visualization and comparison. Figure 6 shows the AUC scores of different machine learning classifiers when trained on a harmonized and unharmonized dataset. Results show that we got better predictive scores than the unharmonized dataset when using harmonization. Figure 7 shows the feature selection's AUC scores of different machine learning classifiers. When harmonization was applied, we observed a similar AUC score increase in machine learning classifiers. Figure 6 and Figure 7 showed that with or without the feature selection, the highest AUC score of 0.85 was achieved for the nearest neighbor classifier for harmonized dataset compared to the highest of 0.76 on the unharmonized dataset. The performance of the algorithms on harmonized datasets changes very slightly in a few cases.

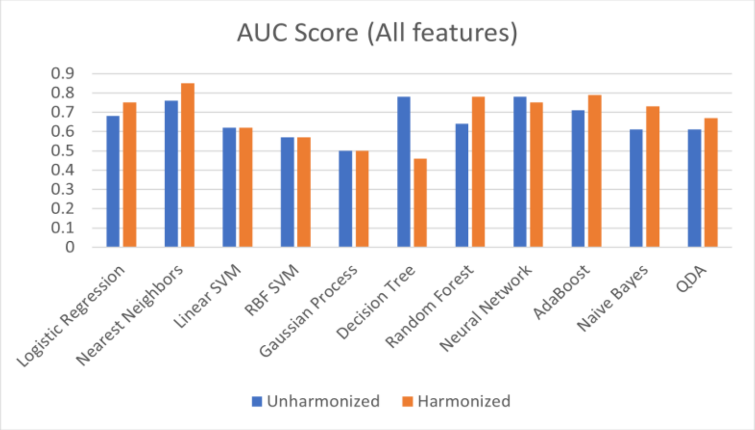

Figure 6 AUC score comparison of different classifier for unharmonized and harmonized datasets considering all features

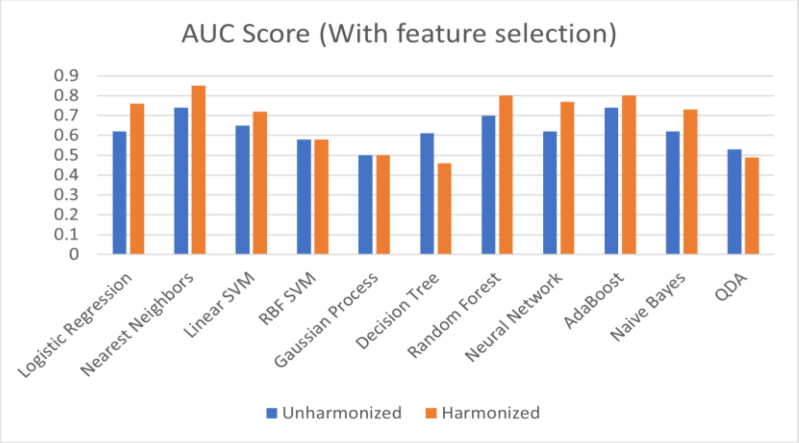

Figure 7 AUC score comparison of different classifier for unharmonized and harmonized datasets considering feature selection.

6 DISCUSSION

In our work, we proposed a scalable decentralized version of ComBat which can be used for harmonizing neuroimage datasets in a decentralized fashion. From the algorithm presented above, we can observe that our model only shares simple meta-information about datasets which helps each site harmonize its dataset independently with respect to other participating sites. Also, no complex operation was performed in the remote node, so it does not require high computational power. This model has several advantages. First, data sharing becomes more manageable as it does not require the dataset transfer away from the original location. Secondly, we do not need to create redundant copies of the datasets by pooling them on a single location, saving much space and reducing the computational cost associated. Thirdly, our model can be easily extended when participating sites increase. Fourthly, each node harmonizes its dataset independently, which requires less computational power. Fifthly, our model is integrated with COINSTAC, providing additional security during information exchange off the shelf. Finally, we can easily combine our model with other decentralized algorithms provided by COINSTAC to create different analysis pipelines. Another contribution of our work is that there is no significant difference between the computer parameters of centralized ComBat and decentralized ComBat.

We presented a simple star network model which could harmonize data in a decentralized environment. Also, from Figure 1, it can be observed that original data never leaves the sites, which protects the confidentiality of the datasets. Also, the computational cost is divided among the local nodes.

We observed the influence of site effects in the dataset before and after harmonization. After harmonization, we observed increased connectivity among the mTBI groups because harmonization removed the site effects. Moreover, results in post-harmonized data Figure 3 suggest that mTBI patients develop hyperconnectivity after TBI injuries. Based on the literature, increased connectivity is a regular observation in TBI as the brain reacts to the traumatic injury event [28,29,30]. In our case, after we removed the site effects from the datasets, we observed more connectivity in the TBI groups not observed before as it was mixed with site effects. Based on the observations, we can say that harmonization does help in removing confounding non-biological effects allowing for more meaningful discoveries.

In our study, we showed that our proposed model could handle an increased number of sites. Based on the simulation, we showed that DC-ComBat could harmonize even 250 sites simultaneously.

We showed in Figure 4 and Figure 5 that after the number of sites reached above 50, there was no change in performance. Moreover, the remote node does not perform any complex operation. Instead, all the complex operation, such as harmonization, is done on each local node. That is why the model can scale quickly when participating nodes increase.

In our study, we observed that the performance of DC-ComBat is dependent on the two site parameters called additive parameter and multiplicative parameter. The base assumption of the ComBat model is that the multiplicate parameter will follow the inverse-gamma distribution, and the additive parameter will follow the Gaussian distribution. However, we cannot control the probability distributions of site parameters directly. That is why our proposed model may perform poorly for some distributions for the Poisson distribution shown in Figure 4 and Figure 5. We observed that Rayleigh and Weibull distributions were similar to the inverse-gamma because they conjugate prior to inverse-gamma [31], whereas Poisson is not for the inverse-gamma distribution. Moreover, we also observed that skewness and kurtosis could increase or decrease the performance of our model. We will not discuss the effects of probability distributions of site effects as it is not fully understood and will be a part of our future research direction.

We also extended our study to observe how machine learning algorithms can perform better on a harmonized dataset. Data collected at two sites that used different scanners, parameters and acquisition methods, machine learning classifiers performed relatively poorly in this study. After including data harmonization in the machine learning pipeline, we found that reducing site effects can improve machine learning classifier performance. The New Mexico dataset was analyzed in a previous classification study where the authors showed that SVM has an AUC score of 0.85 [32]. Another study used SVM to discriminate mTBI patients from healthy controls (HC) on a different dataset and found an AUC score of 0.72 [33]. However, when we combined the two-sites datasets, we found that the performance of the SVM classifier decreased to an AUC score of 0.62. When we harmonized the data and performed feature selection SVM algorithm reached an AUC score of 0.72. The predictive score decreased because the site effect heavily affected the combined dataset. After harmonization, we observed a high correlation between the t-values of group difference and feature importance. Moreover, the previous studies only considered single-source data collected by a single scanner and the same acquisition methods. This contributed to our hypothesis that harmonization improves performance for multi-site analysis.

The main contribution of this work is the decentralization of the harmonization process using ComBat and COINSTAC. The output of these two separate approaches had very insignificant differences due to the difference in machines precision and operating systems. Therefore, we conclude that both approaches produce identical output. Our proposed model is more optimal than the centralized approach considering the volume, confidentiality, security, and resource constraints associated with data.

7 LIMITATIONS AND FUTURE DIRECTIONS

There are several limitations in the current study, which will be addressed in future studies. We did not concern about the re-identification attack; we only secured the intercommunications between local and remote nodes. Our study worked with FNC datasets; however, we could study other image modalities in our subsequent studies. Moreover, we did not present many details related to the site parameter distributions as we had no accurate knowledge about the probability distribution of site parameters to compare. In future studies, we want to add differential privacy and study the effects of site parameter distribution in more detail.

8 CONCLUSION

The proposed novel model showed that decentralized algorithms could achieve identical results as their centralized counterpart. Also, the decentralized approaches solve many challenges associated with data sharing and connecting the whole world. This study encouraged future researchers to contribute to making new decentralized algorithms, which will help us study all the data scattered across the world and produce beneficiary outcomes.

REFERENCES

1. Bassett, Danielle S, and Olaf Sporns. "Network neuroscience." Nature neuroscience vol. 20,3 (2017): 353-364. doi:10.1038/nn.4502
2. Amunts, K., Ebell, C., Muller, J., Telefont, M., Knoll, A., and Lippert, T. (2016). The human brain project: creating a european research infrastructure to decode the human brain. Neuron 92, 574–581. doi: 10.1016/j.neuron.2016.10.046
3. Guger, Christoph, Brendan Z. Allison, and Aysegul Gunduz. "Brain-Computer Interface Research: A State-of-the-Art Summary 10." In Brain-Computer Interface Research, pp. 1-11. Springer, Cham, 2021.
4. Thompson, P. M., Stein, J. L., Medland, S. E., Hibar, D. P., Vasquez, A. A., Renteria, M. E., et al. (2014). The enigma consortium: large-scale collaborative analyses of neuroimaging and genetic data. Brain Imaging Behav. 8, 153–182. doi: 10.1007/s11682-013-9269-5
5. Thompson, P. M., Andreassen, O. A., Arias-Vasquez, A., Bearden, C. E., Boedhoe, P. S., Brouwer, R. M., et al. (2015). ENIGMA and the individual: predicting factors that affect the brain in 35 countries worldwide. NeuroImage. doi: 10.1016/j.neuroimage.2015.11.057.
6. T. Sherif, P. Rioux, M.-E. Rousseau, N. Kassis, N. Beck, R. Adalat, S. Das, T. Glatard, and A. C. Evans. CBRAIN: a web-based, distributed computing platform for collaborative

neuroimaging research. Frontiers in Neuroinformatics, 8(54), May 2014. doi: 10.3389/fninf.2014.00054.

7. Drew Landis, William Courtney, Christopher Dieringer, Ross Kelly, Margaret King, Brittny Miller, Runtang Wang, Dylan Wood, Jessica A. Turner, Vince D. Calhoun, COINS Data Exchange: An open platform for compiling, curating, and disseminating neuroimaging data, NeuroImage,Volume 124, Part B,2016,Pages 1084-1088,ISSN 1053-8119, https://doi.org/10.1016/j.neuroimage.2015.05.049.
8. Glover, Gary H et al. "Function biomedical informatics research network recommendations for prospective multicenter functional MRI studies." Journal of magnetic resonance imaging : JMRI vol. 36,1 (2012): 39-54. doi:10.1002/jmri.23572
9. Jean-Philippe Fortin, Drew Parker, Birkan Tunc, Takanori Watanabe, Mark A Elliott, Kosha Ruparel, David R Roalf, Theodore D Satterthwaite, Ruben C Gur, Raquel E Gur, Robert T Schultz, Ragini Verma, Russell T Shinohara. Harmonization Of Multi-Site Diffusion Tensor Imaging Data. NeuroImage, 161, 149-170, 2017
10. W. Evan Johnson and Cheng Li, Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics, 8(1):118-127, 2007.
11. Jean-Philippe Fortin, Nicholas Cullen, Yvette I. Sheline, Warren D. Taylor, Irem Aselcioglu, Philip A. Cook, Phil Adams, Crystal Cooper, Maurizio Fava, Patrick J. McGrath, Melvin McInnis, Mary L. Phillips, Madhukar H. Trivedi, Myrna M. Weissman, Russell T. Shinohara. Harmonization of cortical thickness measurements across scanners and sites. NeuroImage, 167, 104-120, 2018

12. Yu, M, Linn, KA, Cook, PA, et al. Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. Hum Brain Mapp. 2018; 39: 4213 – 4227. https://doi.org/10.1002/hbm.24241

13. Orlhac, F. et al. A post-reconstruction harmonization method for multicenter radiomic studies in pet. J Nucl Med. https://doi.org/10.2967/jnumed.117.199935 (2018).

14. S. Plis, A.D. Sarwate, D. Wood, C. Dieringer, D. Landis, C. Reed, S.R. Panta, J.A. Turner, J.M. Shoemaker, K.W. Carter, P. Thompson, K. Hutchison, V.D. Calhoun, COINSTAC: A Privacy Enabled Model and Prototype for Leveraging and Processing Decentralized Brain Imaging Data, Frontiers in Neuroscience 10(365): August 2016. http://dx.doi.org/10.3389/fnins.2016.00365

15. B. Baker, A. Abrol, R.F. Silva, E. Damaraju, A.D. Sarwate, V.D. Calhoun, S.M. Plis, Decentralized Temporal Independent Component Analysis: Leveraging fMRI Data in Collaborative Settings, NeuroImage 186: pp. 557–569, February 2019. http://dx.doi.org/10.1016/j.neuroimage.2018.10.072

16. Wojtalewicz, N. P., Silva, R. F., Calhoun, V. D., Sarwate, A. D., and Plis, S. M. (2017). "Decentralized independent vector analysis," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (New Orleans, LA: IEEE), 826–830.

17. Lewis, N., Plis, S., and Calhoun, V. (2017). "Cooperative learning: Decentralized data neural network," in 2017 International Joint Conference on Neural Networks (IJCNN) (Anchorage, AK), 324–331.

18. Baker, B. T., Silva, R. F., Calhoun, V. D., Sarwate, A. D., and Plis, S. M. (2015). "Large scale collaboration with autonomy: Decentralized data ICA," in 2015 IEEE 25th

International Workshop on Machine Learning for Signal Processing (MLSP), (Boston, MA: IEEE), 1–6.

19. H. Imtiaz, A.D. Sarwate, Differentially Private Distributed Principal Component Analysis, Proceedings of the 43rd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, pp. 2206–2210, 15–20 April 2018. http://dx.doi.org/10.1109/ICASSP.2018.8462519
20. Vergara VM, Mayer A, Kiehl KA, Calhoun VD. Dynamic functional network connectivity discriminates mild traumatic brain injury through machine learning. NeuroImage: Clinical 2018
21. Ling JM, Peña A, Yeo RA, Merideth FL, Klimaj S, Gasparovic C, et al. Biomarkers of increased diffusion anisotropy in semi-acute mild traumatic brain injury: a longitudinal perspective. Brain 2012; 135(4): 1281-92
22. Vergara VM, Mayer AR, Damaraju E, Kiehl KA, Calhoun V. Detection of Mild Traumatic Brain Injury by Machine Learning Classification Using Resting State Functional Network Connectivity and Fractional Anisotropy. J Neurotrauma 2017; 34(5): 1045-53.
23. Friston KJ. Statistical parametric mapping. Neuroscience Databases: Springer; 2003. p. 237-50.
24. V.Calhoun, T.Adali, G.Pearlson, and J.Pekar,Group ICA of Functional MRI Data: Separability, Stationarity, and InferenceProceedings, ICA2001, San Diego, CA, 2001.
25. GIFT software (GIFT v4: https://trendscenter.org/software/gift/)
26. M. S. Salman, Y. Du, E. Damaraju, Q. Lin and V. D. Calhoun, "Group information guided ICA shows more sensitivity to group differences than dual-regression," 2017 IEEE 14th

International Symposium on Biomedical Imaging (ISBI 2017), 2017, pp. 362-365, doi: 10.1109/ISBI.2017.7950538.

27. Vergara, V., Weiland, B., Hutchison, K. et al. The Impact of Combinations of Alcohol, Nicotine, and Cannabis on Dynamic Brain Connectivity. Neuropsychopharmacol. 43, 877–890 (2018). https://doi.org/10.1038/npp.2017.280
28. Hillary, Frank G et al. "The rich get richer: brain injury elicits hyperconnectivity in core subnetworks." PloS one vol. 9,8 e104021. 14 Aug. 2014, doi:10.1371/journal.pone.0104021
29. Morelli, Nathan et al. "Resting state functional connectivity responses post-mild traumatic brain injury: a systematic review." Brain injury, 1-12. 6 Sep. 2021, doi:10.1080/02699052.2021.1972339
30. Vergara, Victor M et al. "Detection of Mild Traumatic Brain Injury by Machine Learning Classification Using Resting State Functional Network Connectivity and Fractional Anisotropy." Journal of neurotrauma vol. 34,5 (2017): 1045-1053. doi:10.1089/neu.2016.4526
31. https://en.wikipedia.org/wiki/Conjugate_prior
32. Vergara, V. M., Mayer, A. R., Damaraju, E., Kiehl, K. A., & Calhoun, V. (2017). Detection of Mild Traumatic Brain Injury by Machine Learning Classification Using Resting State Functional Network Connectivity and Fractional Anisotropy. Journal of neurotrauma, 34(5), 1045–1053. https://doi.org/10.1089/neu.2016.4526
33. Xiaoping Luo, Dezhao Lin, Shengwei Xia, Dongyu Wang, Xinmang Weng, Wenming Huang, Hongda Ye, "Machine Learning Classification of Mild Traumatic Brain Injury Using Whole-Brain Functional Activity: A Radiomics Analysis", Disease Markers, vol. 2021, Article ID 3015238, 7 pages, 2021. https://doi.org/10.1155/2021/3015238