A Decentralized ComBat Algorithm and Applications to Functional Network Connectivity

Bostami, Biozid; Hillary, Frank G.; van der Horn, Harm Jan; van der Naalt, Joukje; Calhoun, Vince D.; Vergara, Victor M.

Published in:
Frontiers in Neurology

DOI:
10.3389/fneur.2022.826734

Publication date:
2022

frontiers in Neurology

METHODS
published: 15 March 2022
doi: 10.3389/fneur.2022.826734

A Decentralized ComBat Algorithm and Applications to Functional Network Connectivity

Biozid Bostami[1,2,3], Frank G. Hillary[4], Harm Jan van der Horn[5], Joukje van der Naalt[6], Vince D. Calhoun[1,2,3] and Victor M. Vergara[1,2,3]**

[1] Department of Computer Science, Tri-institutional Center for Translational Research in Neuroimaging and Data Science, Georgia State University, Atlanta, GA, United States, [2] Department of Computer Science, Georgia Institute of Technology, Atlanta, GA, United States, [3] Department of Computer Science, Emory University, Atlanta, GA, United States, [4] Department of Psychology, Penn State University, State College, PA, United States, [5] Department of Developmental Neurology, University of Groningen, University Medical Center Groningen, Groningen, Netherlands, [6] Department of Medical Sciences, University of Groningen, University Medical Center Groningen, Groningen, Netherlands

Recent studies showed that working with neuroimage data collected from different research facilities or locations may incur additional source dependency, affecting the overall statistical power. This problem can be mitigated with data harmonization approaches. Recently, the ComBat method has become commonly adopted for various neuroimage modalities. While open neuroimaging datasets are becoming more common, a substantial amount of data is still unable to 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 that does not create redundant copies of the original datasets and performs remote operations on the datasets separately without sharing any individual subject data, ensuring a certain level of privacy and reducing regulatory hurdles. We proposed a novel approach called "Decentralized ComBat" which can harmonize datasets separately without combining the datasets. We tested our model by harmonizing functional network connectivity datasets from two traumatic brain injury studies in a decentralized way. Also, we used simulations to analyze the performance and scalability of our model when the number of data collection sites increases. We compare the output with centralized ComBat and show that the proposed approach produces similar results, increasing the sensitivity of the functional network connectivity analysis and validating our approach. Simulations show that our model can be easily scaled to many more datasets based on the requirement. In sum, we believe this provides a powerful tool, further complementing open data and allowing for integrating public and private datasets.

Keywords: harmonization, federated learning, neuroimage analysis, functional connectivity, brain network

OPEN ACCESS

Edited by:
Maxime Descoteaux, Université de Sherbrooke, Canada

Reviewed by:
Fabrizio Pizzagalli, University of Turin, Italy
Anderson M. Winkler, National Institute of Mental Health (NIH), United States

****Correspondence:***
Biozid Bostami
bbostami1@student.gsu.edu
Victor M. Vergara
vvergarascience@gmail.com

Specialty section:
This article was submitted to Applied Neuroimaging, a section of the journal Frontiers in Neurology

Received: *01 December 2021*
Accepted: *09 February 2022*
Published: *15 March 2022*

Citation:
Bostami B, Hillary FG, van der Horn HJ, van der Naalt J, Calhoun VD and Vergara VM (2022) A Decentralized ComBat Algorithm and Applications to Functional Network Connectivity. Front. Neurol. 13:826734. doi: 10.3389/fneur.2022.826734

INTRODUCTION

The significance of network neuroscience has reached a global scale with a growing number of large-scale projects related to impactful topics such as brain disease, brain development, brain aging, and brain-computer interfacing (1–3). Maximizing the potential of these large projects to reach their goal 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 can consume a large amount of time. Another significant concern is the volume of the data collected from multiple sites because merging large neuroimage datasets in a single location consumes redundant 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 toward decentralized data-sharing (6, 7) which is particularly true with widespread efforts in the neuroimage 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), Dynamic Functional Network Connectivity (13). However, the current ComBat model does not address data access problems, including geographical and confidentiality issues, which motivate us to develop a decentralized model that works in a distributed environment. This manuscript presents a decentralized harmonization model called "Decentralized ComBat (DC-ComBat)".

For several years, our team has been working on a web-based framework 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 that returns similar results on collected datasets with this framework. 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 computation proposed earlier include 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. We found this framework suitable for our decentralized approach based on the benefits.

METHODS

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

$$y_{i,j,v}=\alpha_v+X_{i,j}\beta_v+\gamma_{i,v}+\delta_{i,v}\varepsilon_{i,j,v} \tag{1}$$

In the above equation $\alpha_{\mathbf{v}}$ represents the overall mean value at feature $\boldsymbol{v}$. $\boldsymbol{X}$ represents the biological variants, β_v represents the regression coefficient for $\mathbf{X}$ at feature $\boldsymbol{v}$. The error term $\boldsymbol{\varepsilon}$ is assumed to follow a Gaussian distribution $\mathbf{N(0,\sigma^2)}$. In Equation (1) $\delta_{\mathbf{i},v}$ and $\gamma_{\mathbf{i},v}$ 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:

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

The model can be divided into three parts. The first part is the standardization of data. The Decentralized regression algorithm available in COINSTAC (14) was used to calculate the initial β-coefficients. We calculated the local mean and local variance based on β-coefficients in later stages. After standardization, every data will have similar overall mean and variance. The following equation calculates the standardization data:

$$Z_{i,j,v}=\frac{y_{ijv}-\hat{\alpha}_v-X_{ij}\hat{\beta}_v}{\hat{\sigma}_v} \tag{3}$$

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

$$\gamma_{i,v} \sim N\left(Y_i, \tau_i^2\right) \text{ and } \delta_{i,v}^2 \sim \text{Inverse Gamma}\left(\lambda_i, \theta_i\right) \tag{4}$$

The hyperparameters $\gamma_{\mathbf{i}}, \tau_{\mathbf{i}}^{\mathbf{2}}, \lambda_{\mathbf{i}}, \theta_{\mathbf{i}}$ 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 $\gamma_{i,v}$, $\delta_{i,v}^2$ can be defined by the posteriors means as followings:

$$\gamma_{i,v}^{*} = \frac{n_i\tau_i^2\hat{\gamma_{i,v}}+\delta_{i,v}^{2*}\overline{\gamma}_i}{n_i\tau_i^2+\delta_{i,v}^{2*}} \text{ and } \delta_{i,v}^{2*} = \frac{\overline{\theta}_i+\frac{1}{2}\sum_j\left(Z_{ijv}-\gamma_{i,v}^{*}\right)^2}{n_i\tau_i^2+\delta_{i,v}^{2*}} \quad (5)$$

Finally, data is adjusted based on the estimated site parameters $\gamma_{\mathbf{i,v}}^{*}$ *and* $\delta_{\mathbf{i,v}}^{\mathbf{2}*}$.

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.

Algorithm:

Step1: Initialize the central node and site nodes.

Step2: Collect the initial summary (number of samples) of the site nodes in the central node.

Step3: Calculate the β coefficient for each site using the decentralized regression approach available in COINSTAC.

Step4: 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.

Step5: compute grand mean and grand variance and update each site node.

Step6: For site node i = 1, 2, 3 …. N do

1. standardized 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.

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 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 weights and summary statistics with the remote node via the secured message-passing mechanism [**Figure 1(1)**]. All intermediate communication is encrypted and sent over TLS (Transport Layer Security) provided by COINSTAC (14). The remote node calculates the grand mean and grand variance by aggregating the regional nodes' values in **Figure 1(2)** and broadcasting the grand mean and grand variance to all local nodes in **Figure 1(3)**. 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(4)**]. 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:

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 the performance of our model when it is applied to FNC data as from the previous study presented in (10), which showed that the ComBat model performs well on 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 increases. We tried to simulate a real-world situation where datasets are located at different locations across the world can be harmonized simultaneously. In the following sections, we will describe how these two sets of datasets were collected and pre-processed.

Dataset 1

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

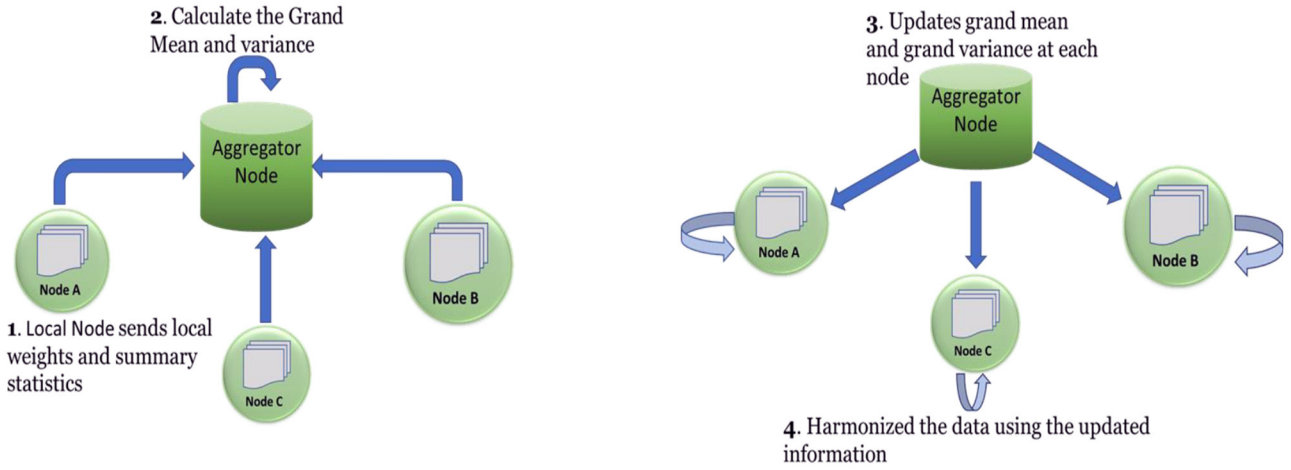

FIGURE 1 | Gives the overall picture of the decentralized ComBat algorithm and intra-communication between nodes.

research publication (22); therefore, we present a brief outline of the whole process.

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) = 2,000 ms; TE (Time of Echo) = 29 ms; flip angle = 75°; FOV (Field of View) = 240 mm; matrix size = 64 × 64. After removing the first five images due to the T1 equilibrium effect, the final 145 images were selected next step analysis.

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 to 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) = 2,000 ms; TE (Time of Echo) = 20 ms; flip angle = 8°; FOV (Field of View) = 224 × 224 × 136.5 mm.

fMRI 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, (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.

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 parameter and Poisson, Rayleigh, or Weibull distribution for multiplicative parameter. 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.

EXPERIMENT SETUP 1 AND INVESTIGATION

For this experiment, we keep two datasets collected from two research facilities into two local nodes. We applied our

model DC-ComBat to harmonize the datasets. To observe the harmonization performance, we perform two different assessments on the dataset. 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 1,128 t-values. Instead of showing vectors, we converted them into a matrix where rows and columns represent each of the 48 ICA components and 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 was adding 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.

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). We observed that 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, found that the maximum difference was $3.06699e^{-15}$. This very slight difference in the output was within the order of magnitude of the machine precision error. We conclude there was no effective difference between ComBat and DC-ComBat.

EXPERIMENT 2 AND INVESTIGATION

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 had 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 had 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 **Figures 4**, **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 **Figures 4**, **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 **Figures 4**, **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 **Figures 4**, **5**, we saw that for Poisson distribution, performance is poor compared to other distributions.

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 the number of participating sites increases. Fourthly, each node harmonizes its dataset independently which requires less computational power. Fifthly, our model is integrated with COINSTAC which provides 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 main 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.

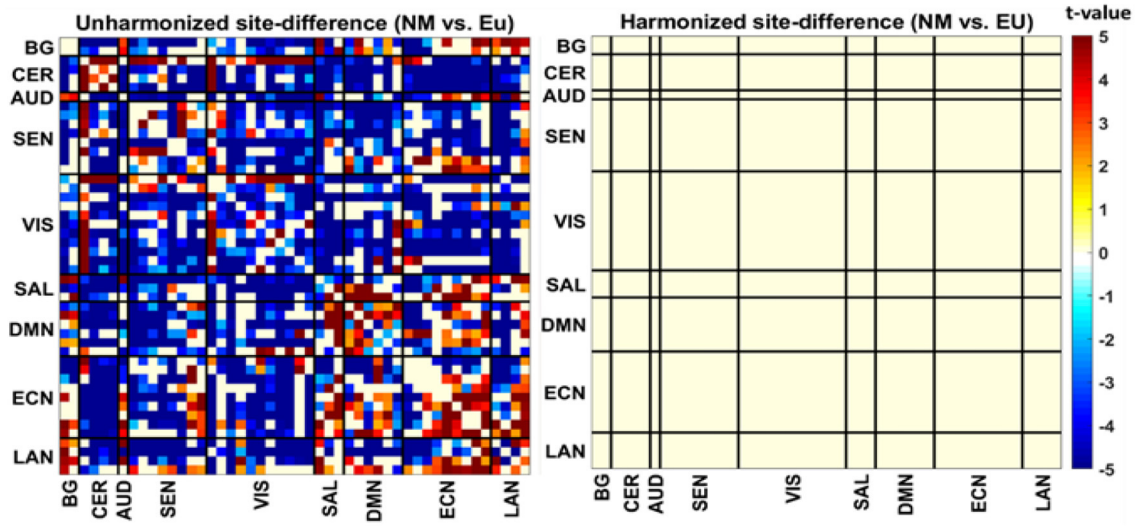

FIGURE 2 | Heatmap of *t*-values site-difference (NM-EU) before **(left)** and after harmonization **(right)**.

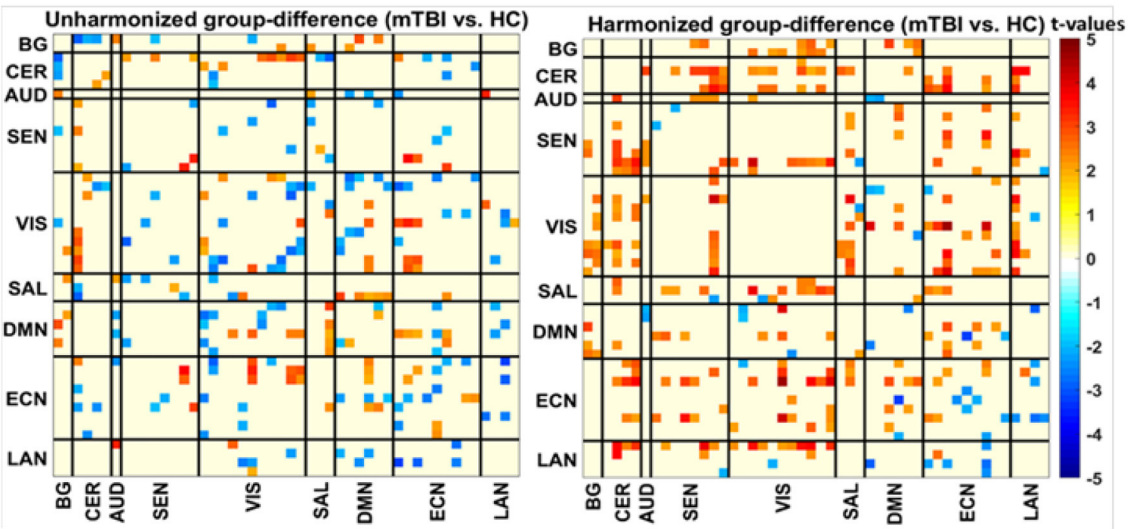

FIGURE 3 | Heatmap of *t*-values group difference (NM-EU) before **(left)** and after harmonization **(right)**.

We observed the influence of site effects in the dataset before and after harmonization. We observed increased connectivity among the mTBI groups after harmonization 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 (22, 28, 29). In our case, after we remove the site effects from the datasets, we observe 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 **Figures 4**, **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 the number of participating nodes increases.

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

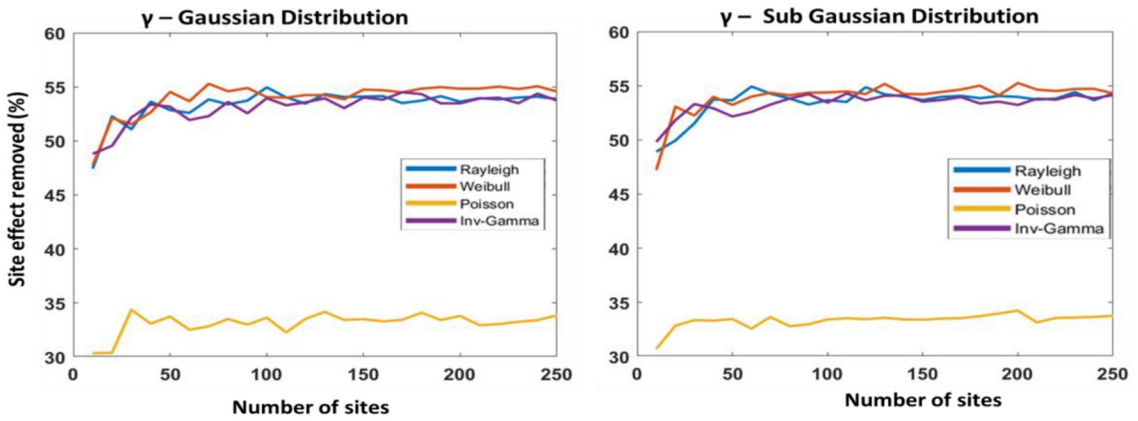

FIGURE 4 | Decentralized ComBat with different distributions as the multiplicative parameter; Gaussian distribution (skewness: 0.26 and kurtosis: 3.3) **(left)** and Sub-Gaussian distribution (skewness: 0.12 and kurtosis: 2.2) **(right)** for additive parameter.

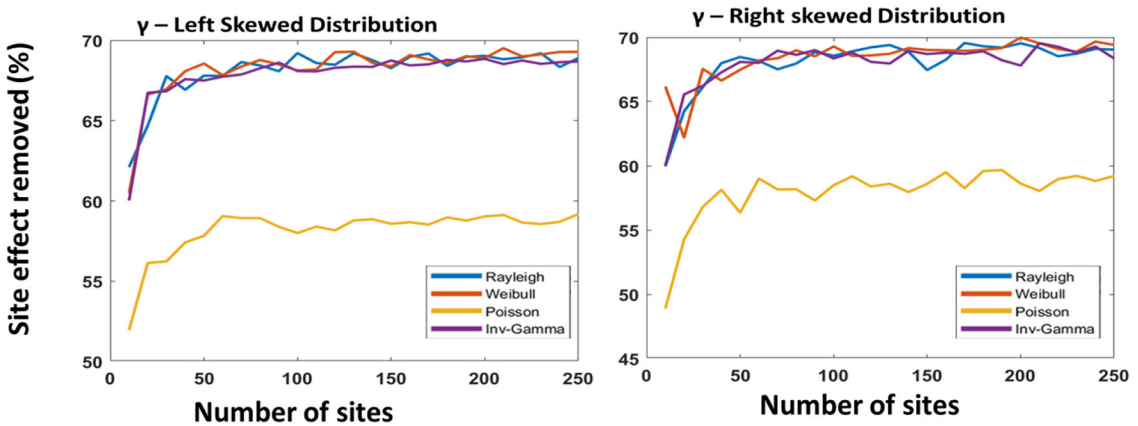

FIGURE 5 | Decentralized ComBat with different distributions as the multiplicative parameter; Skewed-left distribution(skewness: −0.58 and kurtosis: 3.48) **(left)** and right skewed distribution (skewness: 0.34 and kurtosis: 2.34) **(right)** for additive parameter.

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 for some distributions, our proposed model may perform poorly for the Poisson distribution shown in **Figures 4**, **5**. We observed that Rayleigh and Weibull distributions were similar to the inverse-gamma because they conjugate prior to inverse-gamma[1], 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.

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.

LIMITATIONS AND FUTURE DIRECTION

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. We want to add differential privacy and study the effects of site parameter distribution in more detail in our future studies. Moreover, in near future this algorithm will be intrigated with ENIGMA HALFpipe (30).

[1]https://en.wikipedia.org/wiki/Conjugate_prior

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 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.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

BB, VV, VC, and FH: planned the whole project. BB and VV: responsible for conducting full research, writing manuscript, data analysis, and designing the algorithm. VC and VV: result analysis and manuscript revision. VC, VV, and FH: data analysis and result analysis. JN and HH: data collection and processing. All authors contributed to the article and approved the submitted version.

FUNDING

Multiple funds supported this project from the authors VC and FH. We want to thank NIH and NSF for their grants. The grants included in this project are NSF: 2112455 and NIH: R01DA040487 and R61NS120249.

REFERENCES

1. Bassett, Danielle S, Olaf S. Network neuroscience. *Nat Neurosci.* (2017) 20:353–64. doi: 10.1038/nn.4502
2. Amunts K, Ebell C, Muller J, Telefont M, Knoll A, Lippert T. The human brain project: creating a European research infrastructure to decode the human brain. *Neuron*. (2016) 92:574–81. doi: 10.1016/j.neuron.2016.10.046
3. Guger C, Allison BZ, Gunduz A. Brain-computer interface research: a state-of-the-art summary 10. In: Guger C, Allison BZ, Gunduz A, editors. *Brain-Computer Interface Research*. Cham: Springer (2021).
4. Thompson PM, Stein JL, Medland SE, Hibar DP, Vasquez AA, Renteria ME, et al. The enigma consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain Imaging Behav*. (2014) 8:153–82. doi: 10.1007/s11682-013-9269-5
5. Thompson PM, Andreassen OA, Arias-Vasquez A, Bearden CE, Boedhoe PS, Brouwer RM, et al. ENIGMA and the individual: Predicting factors that affect the brain in 35 countries worldwide. *NeuroImage*. (2017) 145:389–408. doi: 10.1016/j.neuroimage.2015.11.057
6. Sherif T, Rioux P, Rousseau M-E, Kassis N, Beck N, Adalat R, et al. CBRAIN: a web-based, distributed computing platform for collaborative neuroimaging research. *Front. Neuroinform.* (2014) 8:54. doi: 10.3389/fninf.2014.00054
7. Landis D, Courtney W, Dieringer C, Kelly R, King M, Miller B, et al. COINS data exchange: an open platform for compiling, curating, and disseminating neuroimaging data. *NeuroImage*. (2016) 124:1084–8. doi: 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. *JMRI.* (2012) 36:39–54. doi: 10.1002/jmri.23572
9. Fortin J-P, Parker D, Tunc B, Watanabe T, Elliott MA, Ruparel K, et al. Harmonization of multi-site diffusion tensor imaging data. *NeuroImage*. (2017) 161:149–70. doi: 10.1016/j.neuroimage.2017.08.047
10. Johnson WE, Li C. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. (2007) 8:118–27. doi: 10.1093/biostatistics/kxj037
11. Fortin J-P, Cullen N, Sheline YI, Taylor WD, Aselcioglu I, Cook PA, et al. Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*. (2018) 167:104–20. doi: 10.1016/j.neuroimage.2017.11.024
12. Yu M, Linn KA, Cook PA, Phillips ML, McInnis M, Fava M, et al. Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. *Hum Brain Mapp.* (2018) 39:4213–27. doi: 10.1002/hbm.24241
13. Bostami B, Calhoun VD, Van Der Horn HJ, Vergara V. Harmonization of multi-site dynamic functional connectivity network data. In: *2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE)* (2021), p. 1–4, doi: 10.1109/BIBE52308.2021.9635538
14. Plis S, Sarwate AD, Wood D, Dieringer C, Landis D, Reed C, et al. COINSTAC: a privacy enabled model and prototype for leveraging and processing decentralized brain imaging data. *Front Neurosci.* (2016) 10:e00365. doi: 10.3389/fnins.2016.00365
15. Baker B, Abrol A, Silva RF, Damaraju E, Sarwate AD, Calhoun VD, et al. Decentralized temporal independent component analysis: leveraging fMRI data in collaborative settings. *NeuroImage.* (2019) 186:557–69. doi: 10.1016/j.neuroimage.2018.10.072
16. Wojtalewicz NP, Silva RF, Calhoun VD, Sarwate AD, Plis SM. Decentralized independent vector analysis. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA: IEEE. (2017). p. 826–30.
17. Lewis N, Plis S, Calhoun V. Cooperative learning: decentralized data neural network. In: *2017 International Joint Conference on Neural Networks (IJCNN)*, Anchorage, AK (2017). p. 324–31.
18. Baker BT, Silva RF, Calhoun VD, Sarwate AD, Plis SM. Large scale collaboration with autonomy: decentralized data ICA. In: *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, Boston, MA: IEEE (2015). p. 1–6.
19. Imtiaz H, Sarwate AD. Differentially private distributed principal component analysis. In: *Proceedings of the 43rd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Calgary, AB (2018). p. 2206–2210, 15–20.
20. Vergara VM, Mayer A, Kiehl KA, Calhoun VD. Dynamic functional network connectivity discriminates mild traumatic brain injury through machine learning. *NeuroImage: Clinical*. (2018) 19:30–7. doi: 10.1016/j.nicl.2018.03.017
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) 135:1281–92. doi: 10.1093/brain/aws073
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) 34:1045–53. doi: 10.1089/neu.2016.4526
23. Friston KJ. Statistical parametric mapping. In: Kötter R, editors. *Neuroscience Databases*. Boston, MA: Springer (2003). doi: 10.1007/978-1-4615-1079-6_16
24. Calhoun V, Adali T, Pearlson G, Pekar J. *Group ICA of Functional MRI Data: Separability, Stationarity, and Inference Proceedings, ICA2001* (2001). San Diego, CA.
25. *GIFT Software.* GIFT v4. Available online at: https://trendscenter.org/software/gift/ (accessed August 11, 2021).
26. Salman MS, Du Y, Damaraju E, Lin Q, Calhoun VD. Group information guided ICA shows more sensitivity to group differences than dual-regression.

In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)* Melbourne, VIC (2017). p. 362–65.
27. Vergara V, Weiland B, Hutchison K, Calhoun V. The impact of combinations of alcohol, nicotine, and cannabis on dynamic brain connectivity. *Neuropsychopharmacol*. (2018) 43:877–90. doi: 10.1038/npp.2017.280
28. Hillary FG, Rajtmajer SM, Roman CA, Medaglia JD, Slocomb-Dluzen JE, Calhoun VD, et al. The rich get richer: brain injury elicits hyperconnectivity in core subnetworks. *PLoS ONE*. (2014) 9:e104021. doi: 10.1371/journal.pone.0104021
29. Morelli N, Johnson NF, Kaiser K, Andreatta RD, Heebner NR, Hoch MC. Resting state functional connectivity responses post-mild traumatic brain injury: a systematic review. *Brain Inj*. (2021) 35:1326–37. doi: 10.1080/02699052.2021.1972339
30. Waller L, Erk S, Pozzi E, Toenders YJ, Haswell CC, Büttner M, et al. ENIGMA HALFpipe: Interactive, reproducible, and efficient analysis for resting-state and task-based fMRI data. *bioRxiv [Preprint]*. (2021). doi: 10.1101/2021.05.07.442790

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.