HHS Public Access

Author manuscript

Inform Med Unlocked. Author manuscript; available in PMC 2023 April 06.

Published in final edited form as:

Inform Med Unlocked. 2023 ; 37: . doi:10.1016/j.imu.2023.101176.

Towards greater neuroimaging classification transparency via the integration of explainability methods and confidence estimation approaches

Charles A. Ellis[a,b,*], **Robyn L. Miller**[b,c], **Vince D. Calhoun**[a,b,c]

[a]Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, 313 Ferst Dr NW, Atlanta, 30332, GA, United States

[b]Tri-institutional Center for Translational Research in Neuroimaging and Data Science: Georgia State University, Georgia Institute of Technology, Emory University, 55 Park Pl NE, Atlanta, GA, 30303, United States

[c]Department of Computer Science, Georgia State University, 25 Park PlaceSuite 700, Atlanta, GA, 30303, United States

Abstract

The field of neuroimaging has increasingly sought to develop artificial intelligence-based models for neurological and neuropsychiatric disorder automated diagnosis and clinical decision support. However, if these models are to be implemented in a clinical setting, transparency will be vital. Two aspects of transparency are (1) confidence estimation and (2) explainability. Confidence estimation approaches indicate confidence in individual predictions. Explainability methods give insight into the importance of features to model predictions. In this study, we integrate confidence estimation and explainability approaches for the first time. We demonstrate their viability for schizophrenia diagnosis using resting state functional magnetic resonance imaging (rs-fMRI) dynamic functional network connectivity (dFNC) data. We compare two confidence estimation approaches: Monte Carlo dropout (MCD) and MC batch normalization (MCBN). We combine them with two gradient-based explainability approaches, saliency and layer-wise relevance propagation (LRP), and examine their effects upon explanations. We find that MCD often adversely affects model gradients, making it ill-suited for integration with gradient-based explainability methods. In contrast, MCBN does not affect model gradients. Additionally, we find many participant-level differences between regular explanations and the distributions of explanations for combined explainability and confidence estimation approaches. This suggests that a similar confidence estimation approach used in a clinical context with explanations only output for the regular model would likely not yield adequate explanations. We hope that our findings will provide a starting point for the integration of the two fields, provide useful guidance for future

*Corresponding author. Tri-institutional Center for Translational Research in Neuroimaging and Data Science: Georgia State University, Georgia Institute of Technology, Emory University, 55 Park Pl NE, Atlanta, GA, 30303, United States. cae67@gatech.edu (C.A. Ellis).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

studies, and accelerate the development of transparent neuroimaging clinical decision support systems.

Keywords

Neuroimaging; Explainable artificial intelligence; Gradient-based explainability methods; Monte Carlo dropout; Monte Carlo batch normalization; Clinical decision support systems

1. Introduction

In recent years, studies have increasingly sought to develop automated diagnosis approaches using machine learning and deep learning methods for a variety of neurological and neuropsychiatric disorders like schizophrenia [1–3], major depressive disorder [4,5], Alzheimer's disease [6,7], and others. This growth can be partially attributed to the limitations of existing clinical diagnostic approaches that are often dependent solely upon symptoms, rather than empirical biological markers, for diagnosis [8–10]. As many disorders can have overlapping symptoms, this is particularly problematic and can lead to delays in diagnoses and misdiagnoses. Nevertheless, if the methods being developed for automated diagnosis are ever to be implemented in a clinical setting, model transparency must be taken into consideration [11]. While there are multiple dimensions to transparency [12], there is little or no available literature on the subject of integrating two of those dimensions - model confidence estimates [13,14] and model explainability [2,15] – into the same model for neuroimaging classification. In this study, we compare two existing approaches for estimating model confidence [13,14]. We then examine their compatibility with two popular gradient-based explainability approaches [16,17] within the context of resting state functional magnetic resonance imaging (rs-fMRI) classification of schizophrenia (SZ), providing guidance and a starting point for future studies seeking to develop more transparent neuroimaging models.

Within the context of automated diagnosis of neurological and neuropsychiatric disorders, and SZ in particular, a variety of modalities like magnetoencephalography (MEG) [10], electroencephalography (EEG) [8,9,18], magnetic resonance imaging (MRI) [19–21], and functional magnetic resonance imaging (fMRI) [2,6,22] have been used. Relative to MRI, fMRI offers insight into the links between schizophrenia and brain dynamics. Relative to modalities like MEG and EEG, fMRI is recorded at lower sampling rates and thus provides less insight into the activity of brain dynamics. However, fMRI offers significantly enhanced spatial resolution and localization relative to MEG and EEG. Additionally, fMRI has been used in many studies related to schizophrenia [23–26]. As such, it represents a useful modality for eventual use within the context of automated diagnosis and clinical decision support. Within the context of rs-fMRI analysis, many studies have utilized functional network connectivity (FNC) for insight into the interaction of brain networks [26–30], so FNC represents a useful starting point for the automated diagnosis of SZ [26].

While the use of rs-fMRI FNC for automated diagnosis of neurological disorders like SZ represents a significant opportunity within the context of clinical decision support [31], black-box automated clinical decision support systems are unlikely to be accepted by

clinicians. As such, model transparency represents a critical component of the eventual implementation of clinical decision support systems. An important aspect of transparency is the capacity to provide an estimate of confidence in predictions [12]. To this end, one approach, called Monte Carlo dropout (MCD), has seen increasing use within the domain of neuroimaging classification [13]. MCD has been used in a variety of studies, including those focused on cortex parcellation [32], dynamics estimation [33,34], and classification of autism spectrum disorder [35] and Parkinson's disease [36]. A more recently developed alternative to MCD that has seen comparatively little use in the domain of neuroimaging classification is Monte Carlo batch normalization (MCBN) [14]. A comparison of the two methods for the domain of neuroimaging classification could provide a useful point of reference for future studies.

The use of methods like MCD and MCBN that can provide estimates of prediction confidence will greatly assist the development of transparent clinical decision support systems. However, in and of themselves, they are insufficient to the task. It is also critical that automated neuroimaging-based clinical decision support systems be explainable [11,37–39]. If clinicians are to use clinical decision support systems, they are ethically obligated to be able to explain the recommendations of such systems to their patients [11]. Both explainability methods [2,40,41] and more recently developed interpretable models [42–44] have been used extensively within the domain of neuroimaging analysis. Nevertheless, with the exception of our preliminary work on this topic [1], explainability methods and approaches for estimating model confidence have, to our knowledge, not yet been integrated.

As both approaches are necessary for the long-term development of clinical decision support systems, the lack of integration of confidence estimation approaches and explainability methods remains a key gap in the current capabilities of the field. In this study, we compare MCD and the more recent MCBN approaches for estimating classifier confidence to better understand their relative utility for the domain of neuroimaging classification. We then, for the first time, integrate two popular gradient-based explainability methods with MCD and MCBN, evaluating the effects of MCD and MCBN upon the explanations and seeking to determine the best approach for integrating the two transparency approaches. It is our hope that this study will provide guidance and a helpful starting point to future studies seeking to develop more transparent neuroimaging models and clinical decision support systems.

2. Methods

2.1. Data collection

We used the Functional Imaging Biomedical Informatics Research Network (FBIRN) dataset consisting of rs-fMRI data from 151 individuals with SZ and 160 healthy controls (HCs). The dataset has been in a number of previous studies [25,26]. The data was collected from the University of California at Irvine, the University of California at Los Angeles, the University of North Carolina at Chapel Hill, the University of New Mexico, the University of Iowa, and the University of Minnesota. Data collection was approved by the Internal Review Boards of each study site. Six sites used 3T Siemens scanners, and 1 site used a 3T General Electric scanner for collection. T2*-weighted functional images were collected with an AC-PC aligned echo-planar imaging (EPI) sequence (TR = 2s, TE = 30 ms, flip angle =

77°, voxel size = 3.4×3.4 × 4mm^3, slice gap = 1 mm, 162 frames, 5:24 min). Recordings were performed while the participants' eyes were closed.

2.2. Data preprocessing

After collecting the data, we used statistical parametric mapping for preprocessing and used rigid body motion correction to account for head movement. We spatially normalized the recordings to an EPI template in the standard Montreal Neurological Institute (MNI) space, resampling to 3 × 3 × 3mm^3. Lastly, we used a Gaussian kernel with a full width at half maximum of 6 mm to smooth the recordings.

After completing general preprocessing steps, we began the feature extraction process. We applied the Neuromark automated independent component analysis (ICA) pipeline [45] of the GIFT toolbox (http://trendscenter.org/software/gift) using the Neuromark_fMRI_1.0 template (also available at http://trendscenter.org/data) to extract 53 components (ICs) associated with different brain regions and structures. The Neuromark pipeline has been used extensively across a variety of studies within the field [26,28,29]. We then assigned the components to 7 networks – the cerebellar (CBN), default mode (DMN), cognitive control (CCN), visual (VSN), sensorimotor (SMN), auditory (ADN), and subcortical (SCN). After extracting the ICs, we extracted the dFNC values using a sliding window approach. We used a tapered window created by convolving a rectangle of window size = 40s with a Gaussian ($\sigma = 3$). We calculated Pearson's correlation between each of the 53 ICs at each time step, resulting in 1378 dFNC features and 124 time steps per study participant. It should be noted that each of the samples can be assigned to 1 of 28 domain pairs (e.g., connectivity between CCN and VSN is CCN/VSN). After extracting each of the dFNC features, we performed feature-wise z-scoring across all subjects.

2.3. Model development

As shown in Fig. 1, we developed a 1D-CNN architecture with input dimensions of 1378 dFNC features x 124 time steps and output dimensions of 2 (i.e., one probability per class). We applied a 10-fold stratified shuffle split cross-validation approach from Scikit-learn with an approximately 80–10-10 training-validation-test split. Within each fold, we applied data augmentation to triple the number of training samples. This involved adding Gaussian noise ($\mu = 0$, $\sigma = 0.7$) to two copies of each training sample. We used an Adam optimizer with an adaptive learning rate [46]. The learning rate started at 0.001 and decreased by 50% after each 15 epochs without an increase in validation accuracy (ACC). We used Kaiming He normal initialization. To address class imbalances, we used a class-weighted categorical cross-entropy. We trained the model for 100 epochs, shuffling after each epoch and using a batch size of 50. We also used a model checkpoint approach to select the model from the epoch with the highest validation accuracy. When assessing test performance, we calculated the sensitivity (SENS), specificity (SPEC), and ACC for each fold. These metrics are shown in equations (1)–(3), respectively.

$$SENS = \frac{TP}{TP + \mathrm{FN}} \quad \text{Equation 1}$$

$$SPEC = \frac{TN}{FP + TN} \quad \text{Equation 2}$$

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{Equation 3}$$

Where true positives are abbreviated, “TP”, false negatives are abbreviated, “FN”, true negatives are abbreviated, “TN”, and false positives are abbreviated, “FP”.

2.4. Monte Carlo Dropout

Monte Carlo Dropout (MCD) was first presented in Ref. [13]. At a high level, it involves the repeated reinitialization of dropout layers during testing to form a distribution of models, and thus a distribution of test predictions. It hinges upon the realization that dropout can be used to form a Bayesian approximation. In our implementation, we used 1000 iterations of dropout during testing to form a distribution of predictions for each sample.

2.5. Monte Carlo Batch Normalization

Similar to MCD, Monte Carlo Batch Normalization (MCBN) hinges upon the realization that another component of neural networks (i.e., batch normalizations layers) can be used to generate a Gaussian distribution of predictions. MCBN was first presented in Ref. [14]. MCBN involves several steps: (1) randomly selecting a minibatch of training data, (2) updating the model batch normalization layers based upon the minibatch of training data, (3) using the updated model to predict class probabilities for the test data, and (4) repeating steps 1 through 3 a number of times to form a distribution of predictions for each sample. We repeated MCBN for 1000 iterations.

2.6. Explainability

In this study, we applied two gradient-based explainability approaches: saliency and layer-wise relevance propagation.

2.6.1. Saliency—Saliency was one of the first gradient-based explainability methods [17]. It is a fairly straightforward approach that involves taking the gradient of the predicted probability of a sample belonging to a particular class with respect to each of the input features. It indicates the effect that a small change in an input feature has upon the output probability of belonging to a particular class. Larger sensitivity values correspond to a greater level of importance. Saliency has been applied in both neuroimaging studies [47] and studies involving other healthcare data types [48,49].

We applied saliency to both the original network and to the network following each iteration of MCBN and MCD, generating a distribution of values for each feature. Afterwards, we normalized the absolute value of the saliency for each study participant to make sure that the saliency summed to 100.

2.6.2. Layer-wise relevance propagation—Layer-wise relevance propagation (LRP) [50] is a popular gradient-based feature attribution explainability method [51]. It has

been shown to produce less noisy explanations than saliency [52]. It was first developed within the context of image classification. However, because it is widely applicable to a number of deep learning architectures, it has since been applied to a number of data types, including various neuroimaging modalities [40,53] and other healthcare data involving both time-series [54,55] and extracted features [56].

LRP involves a series of steps. (1) A sample is passed through a network and assigned to a particular class. (2) A total relevance value of 1 is assigned to the output node corresponding to the class of the sample. (3) The relevance is propagated back through the network from layer to layer using a relevance rule until the relevance is distributed across the input space. It should be noted that depending upon the relevance rule, there can be both positive and negative relevance. Positive relevance indicates that a particular input feature provides evidence for the sample being assigned to the class to which it was assigned. Negative relevance indicates that a particular input feature gives evidence for a sample being assigned to a class other than that to which it was assigned. In this study, we used the $\alpha\beta$-rule to propagate only positive relevance ($\alpha = 1$, $\beta = 0$). The $\alpha\beta$-rule is shown in equation (4).

$$R_j = \sum_k \left(\alpha \frac{(a_j w_{jk})^+}{\sum_{0,j} (a_j w_{jk})^+} - \beta \frac{(a_j w_{jk})^-}{\sum_{0,j} (a_j w_{jk})^-} \right) R_k \qquad \text{Equation 4}$$

Where the relevance is split into a positive component with a coefficient α and into a negative component with a coefficientβ. The variables k and j indicate a node in a deeper and shallower layer, respectively. The variables a and w indicate the activation associated with a particular node and the weight connecting two nodes in different layers, respectively.

We applied LRP to both the original network and to the network following each iteration of MCBN and MCD, generating a distribution of relevance values for each feature. After propagating relevance through the network in our study, we normalized the absolute relevance for each study participant to make sure that the relevance summed to 100.

2.7. Statistical analyses

We performed six sets of statistical analyses. The first pair of pair of analyses were performed to gain insight into the effects of MCBN and MCD on model predictions. The second pair of analyses were performed to gain insight into the explanations spatially, and the second pair of analyses sought insight into the temporal distribution of importance.

2.7.1. Prediction analyses—To analyze the effects of MCBN and MCD upon model predictions, we performed two analyses. The first analysis sought to understand whether MCBN and MCD mean predictions were, in general, different from the predictions of the basic model. To that end, we performed two non-parametric, two-sided, two-tailed paired Wilcoxon signed-rank tests comparing whether the predicted probability across MCBN or MCD iterations for each test participant across folds was greater than the basic model predicted probability for each test participant across folds. The second analysis sought to understand whether MCD or MCBN moved samples closer to the decision boundary (i.e., 50%). We calculated the absolute difference between the 50% decision boundary and the predicted probabilities for the basic model, the mean predicted probabilities for the MCD

model, and the mean predicted probabilities for the MCBN model for all test participants across folds. We then performed non-parametric, two-sided, one-tailed, paired Wilcoxon signed-rank tests to see if the absolute difference between the mean predicted probabilities and the decision boundary was less for MCD versus the basic model and for MCBN versus the basic model. If one of the tests yielded an insignificant p-value, we performed a follow up test to see if the distance from the boundary was greater.

2.7.1.1. Spatial analyses.: We first sought to understand whether there were differences in the importance of individual network pairs between HCs and SZs and then sought to understand whether there were differences in the spatial distributions of importance between the basic model and the model with MCBN and MCD. (1) For insight into the spatial effects of MCBN and MCD upon the explanations for HCs relative to SZs, we summed the total importance (i.e., normalized relevance for LRP and normalized saliency for saliency) of each dFNC feature across time for each participant. We then averaged the importance within each network domain pair and across participants on a per-fold basis. We then used paired t-tests to determine whether there were differences in spatial importance distributions for HCs versus SZs across folds. We then applied FDR correction ($p < 0.05$). (2) For insight into the how representative the importance values associated with the regular model were of the importance distributions associated with the models with MCBN and MCD, we again summed the total importance of each dFNC feature across time for each subject. We then performed a one-sample *t*-test for each participant and calculated the percentage of participants for which there was a significant difference between the importance values associated with the regular model and the importance distributions of the model with MCBN and MCD. We then applied FDR correction ($p < 0.05$) to the values for each subject on a per-fold and per-feature basis and calculated the mean percentage of participants across folds with significant differences. We implemented these analyses for both LRP and saliency.

2.7.2. Temporal analyses—In our temporal analyses, we first sought to understand whether there were differences in the temporal distribution of importance over time between SZs and HCs and next sought to understand whether there were differences in the temporal distribution of importance between the basic model and model with MCBN and MCD. To this end, we adapted a method presented in Ref. [15]. We used Earth mover's distance (EMD), a distance measure comparing two densities, to calculate the distance for each participant between the importance of each dFNC feature over time with the average importance across time. A smaller EMD value indicates that the importance values are more evenly distributed across time, and a larger EMD value indicates that the importance values are more concentrated within smaller time windows. In our temporal analyses, we summed the total absolute importance for each feature on a per-participant basis and normalized that value such that the total importance summed to 100. We then compared that distribution to a distribution in which the total importance was distributed evenly over time. (1) We first sought to determine whether there were significant differences in the temporal distribution of importance across classes. We calculated the mean EMD values in each network domain pair across HCs and across SZs on a per-fold basis fold and then performed a paired *t*-test comparing the values for HCs and SZs. (2) We next sought to determine whether the MCD

and MCBN importance distributions over time were significantly different from those of the regular model. To do this, we performed one-sample t-tests comparing the MCD and MCBN importance distributions for each dFNC feature with the importance value for the regular model. We next applied FDR correction ($p < 0.05$) to the values for each participant on a per-fold and per-feature basis and calculated the mean percent of participants with significant differences across folds. We repeated these analyses for both LRP and saliency.

3. Analysis of not-a-number (NaN) counts

While our spatial and temporal analyses did provide insight into the distribution of importance values. They overlooked an important aspect of integrating confidence estimate approaches with gradient-based explainability methods. We thought it possible that changes in model gradients associated with MCD might adversely affect the capacity of gradients to be calculated for saliency or for relevance to be propagated for LRP. To this end, we performed two analyses. (1) We calculated the percentage of samples for each fold that returned Not-a-Number (NaN) importance values for the regular model and for at least one of the 100 iterations of MCD and MCBN. (2) We also calculated the percentage of MCD and MCBN iterations for each sample that produced NaN values across folds. In the case of LRP, relevance values might ordinarily be returned in a minority of cases if the total positive and negative relevance for a particular layer summed to zero, while saliency should not ordinarily return NaN values. However, in our implementation of LRP with the $\alpha\beta$-rule, positive and negative relevance should not cancel out because only positive relevance is propagated and negative relevance is effectively assigned a value of zero. The presence of a large percentage of NaN values for MCD and MCBN values would indicate that the methods disrupted model gradients to the extent that the explainability methods became inoperable, which would indicate that they are in some cases incompatible with gradient-based explainability methods. This could foreseeably occur with MCD if enough neurons in successive layers were disabled in a configuration such that there was no path to propagate importance from layer to layer through the network.

4. Results

In this section, we describe the results of our examination of the effects of MCD and MCBN upon model predictions and performance.

4.1. Model performance

Table 1 shows the mean and standard deviation of the performance of our architecture across 10 folds without a confidence estimation (i.e., “Regular”) and with MCD and MCBN. All metrics had mean levels of performance across folds that were much higher than chance-level, with average accuracies around 75%. Interestingly, the model sensitivity and specificity were most balanced for the regular model. In contrast, MCBN and MCD seemed to favor either sensitivity or specificity. MCBN favored SENS much more than it favored SPEC, while MCD favored SPEC much more than it favored SENS. Nevertheless, mean ACC for the model with MCBN was similar to that of the regular model, and the mean ACC was slightly lower for the model with MCD.

4.2. Distribution of predictions

Our two-sided, two-tailed Wilcoxon sign-rank tests found that the mean predictions for MCBN and MCD were significantly different from the basic model ($p < 0.001$). Fig. 2 shows the distribution of sample predictions across folds for the regular model without confidence estimation and for the model with MCD and MCBN. The classifier generally predicted either very high or very low probabilities, even in some cases of misclassification. However, particularly among SZs, there were some samples that fell along a gradient of predictions closer to the 50% line. The range of MCD sample predictions was much wider than the range of MCBN sample predictions. With the exception of samples that were already near the decision boundary, the standard deviation of the predictions for MCBN was typically very small. Lastly, in a large number of instances, MCD seemed to move the sample predictions closer to the 50% decision boundary. This observation was confirmed by our two- sided, one-tailed Wilcoxon sign-rank tests that found that mean MCD predictions were much closer to the 50% decision boundary than the basic model predictions ($p < 0.001$). In contrast, the mean MCBN predictions were farther from the 50% decision boundary than the basic model predictions ($p < 0.001$).

4.3. Spatial analysis

Figs. 3 and 4 show the spatial distributions of importance for LRP and saliency, respectively (i.e., the sum of the absolute value of the importance across all time steps). For both LRP and saliency, there were several brain network pairs with high levels of importance across nodes for the classification, including CBN/SCN and SMN/SCN. Other network pairs, including VSN/SMN, VSN/SCN, and VSN, had high levels of importance for specific brain regions that they contained. Specifically, (1) the interactions of the VSN with the postcentral gyrus of the SMN, (2) the interactions of the VSN with the thalamus of the SCN, and (3) the interactions of the other VSN brain regions with the cuneus region were important. For LRP, the interactions of the SMN with the inferior parietal lobule of the CCN were important. While saliency did not identify this particular interaction as being as important as LRP, it indicated that the interaction of two SMN brain regions with the inferior parietal lobule of the CCN was importance.

Additionally, while there were no significant class-specific differences in saliency values, there were some visible differences in LRP relevance between HCs and SZs. These visible differences were present in the SCN/CBN, the SCN/SCN, the SCN/SMN, the VSN/SMN, the SMN/CBN, the CCN/CBN, and the CBN/CBN. Our statistical analysis of class-specific importance found differences in LRP relevance between HCs and SZs for CCN/SCN and CBN/SMN for both the regular model and model with MCBN. In contrast, there were more network pairs with class- specific relevance differences in models with MCD – VSN/VSN, CCN/ SCN, CCN/CCN, CBN/SMN, and CBN/CBN.

For both LRP and saliency, the overall mean importance across folds seemed to be similar across the regular model, the model with MCD, and the model with MCBN. This indicates that at a high level the three methods yield similar levels of importance. However, Fig. 5 provides some higher resolution insight. It shows the average percentage of samples per fold for which there were significant differences ($p < 0.05$) between the model with regular LRP

and the model with MCBN and MCD. Importantly, for saliency, the majority of samples had regular importance values that were away from the means of the importance values with MCD and with MCBN. Importantly, there were some differences in the percentages of samples for regular versus MCD and for regular versus MCBN in a number of network pairs. Additionally, for LRP, a large number of samples in each fold had significant differences. This was particularly the case for the regular LRP values versus the MCD LRP values, and it was true to a lesser extent in the case for the regular LRP values versus the MCBN LRP values. It should be noted, however, that the use of MCD and MCBN seemed to have a larger effect upon the saliency values than upon the LRP values.

4.4. Temporal analysis

Figs. 6 and 7 show the mean EMD across folds for LRP relevance and saliency, respectively. Higher EMD values indicate that importance is more concentrated temporally, while lower EMD values indicate that importance is more uniformly dispersed over time. Interestingly, for both LRP and saliency, importance is more concentrated temporally in HCs than in SZs across the regular, MCD, and MCBN implementations. This is somewhat unexpected for saliency given that the spatial analyses did not find class-specific differences in importance. Additionally, it is odd that the MCD relevance values tend to be more concentrated temporally than those for the regular and MCBN implementations, while the MCD saliency values seem to be less concentrated temporally than those of the regular and MCBN implementations. The regular and MCBN implementations generally seem to have similar values across both LRP and saliency. It should be noted that while these differences are distinct visually, they are not statistically significant. This indicates that while the differences in the mean EMD can be visualized, the distribution of EMD values do overlap to a degree.

Fig. 8 shows the percent of samples per fold with significant difference between the EMD of the importance for the regular model and with the MCD and MCBN models. Interestingly, for both LRP and saliency, the MCBN importance distributions were further from the importance values of the regular model than the MCD importance values. Additionally, the number of participants with differences was greater for saliency than for LRP.

4.5. NaN value analysis

Lastly, it is important to consider the effects of the confidence estimation approaches upon the viability of the explainability methods. As such, we examined the number of samples and iterations per sample of MCD and MCBN that resulted in the production of NaN values. These results are shown in Fig. 9. Interestingly, saliency with the regular model and with MCBN did not return any NaN values, while LRP with the regular model and with MCBN did on one occasion in two of ten folds. For a significant portion of samples, the model with MCD returned at least one iteration of NaN values for both explainability methods. The percentage of NaN values ranged from around zero to ten percent of iterations for most folds across both saliency and LRP, with some folds reaching much higher levels (e.g., up to 50% for LRP). It should be noted that when a sample returned an explanation with NaN values, those NaN values were returned for all dFNC features and time steps.

5. Discussion

In this section, we discuss the performance results of our model and the results of the effects of MCD and MCBN upon the model predictions. We then discuss the effects of the respective methods upon output explanations. Lastly, we discuss limitations and next steps related to our work.

5.1. Model performance

Overall model performance was well above chance-level. Additionally, while MCBN and MCD were able to provide distributions of predictions for each sample, they seemed to destabilize the SENS and SPEC of the model. This was a surprising finding relative to previous studies that have used MCD and MCBN [57], suggesting that the utility of the methods for improving classifier performance may be somewhat dependent upon the utilized model and dataset. Relative to the performance of the models developed in Ref. [2] that had accuracies ranging from 50% to 83%, our model performance was on par to slightly lower. This could partially be related to the significantly larger dataset size used in Ref. [2]. Additionally, our model performance was below that of [47], which used a novel form of extracted features.

5.2. Distribution of predictions

Our model generally had somewhat polarized predictions for each class, with extremely high probabilities of samples belonging to one class and extremely low probabilities of samples belonging to the other class. While this was the case, there were a number of samples that were misclassified or that were predicted to be closer to the decision boundary line. The predictions of MCBN tended to be more stable across iterations and to move the direction of the predictions further towards the extreme that they already favored. In contrast, the predictions of MCD generally varied widely and moved samples closer to the decision boundary. Thus, it seems that for our data and model the use of dropout during testing more greatly affected predictions than changes in the batch normalization layer values. The comparatively small distribution of MCBN predictions could also be a symptom of the dataset that we employed.

5.3. Spatial analysis

We identified a number of brain network pairs useful to the classification of SZ and HCs. Previous studies have found widespread effects of SZ upon the CBN [2,15,27], SMN [15], and SCN [2] similar to our results. Additionally, some studies have identified differences in the VSN/VSN [25,58], VSN/SCN [25,59], and VSN/SMN [25,60] as important to differentiating SZ from healthy individuals. Specifically, changes in connectivity between the VSN and thalamus of the SCN have been identified [59]. These findings support the overall reliability of our model relative to previous studies.

Our findings on the differences between class-specific LRP and saliency results were largely unsurprising. Given the characteristics of each of the methods. LRP is able to provide class-specific relevance values when using the $\alpha\beta$-rule, so we would expect to see differences in the relevance distribution between classes. In contrast, saliency just shows the gradient

associated with the classification, and it is likely that specific regions will have similar gradients across classes.

Differences between the explanations of the basic model and those of MCD and MCBN can be attributed to the effects that they had upon the model shown in Fig. 2. Namely, MCD had a much greater effect than MCBN upon the model predictions. This greater effect upon predictions implies a greater effect upon the underlying model structure and upon any resulting explanations.

5.4. Temporal analysis

We found that HCs generally had more temporally concentrated importance than SZs. This indicates that the aberrant effects of SZ upon brain network interactions tend to be temporally distributed. This finding contrasts [15], which found effects of SZ upon attention values of a long short-term memory network to be more temporally localized. It should be noted, however, that [15] analyzed the temporal distribution of importance for independent components rather than dFNC. It could indicate that there are temporally localized patterns of brain interaction found in HCs that are disrupted in SZs. Additionally, this finding is related to those of other studies that have found effects of SZ upon brain network dynamics [24,26].

At a high level, the temporal distribution of importance for the regular model and the model with MCBN tended to be more similar to one another than the temporal distribution of the model with MCD. However, on a more fine-grained, per-participant basis, there tended to be a difference between the EMD values associated with the regular model and the EMD distributions for both the model with MCD and with MCBN. Additionally, these differences were stronger for saliency than for LRP.

5.5. NaN value analysis

While MCD did not result in NaN-valued explanations in the majority of iterations, it resulted in NaN-valued explanations much more frequently than MCBN. When the explainability methods output NaN values, they output NaN values for all features, which prevented learning anything about the relative importance of the features. In that regard, the explainability analysis failed, and although the production of NaN-valued explanations can be tied to the use of MCD, the exact reason why NaN values occurred is not completely clear. It is feasible that if enough neurons in subsequent layers were disabled via MCD, there might not be a way to propagate relevance or backpropagation a gradient to the input of a network, or the denominator of Equation (4) may sum to a total relevance of zero. As such, although MCD can sometimes improve model performance, the use of multiple dropout layers can, under some circumstances, disable model gradients, making MCD incompatible with gradient-based explainability methods. This finding could also explain the relatively high variance of model predictions that we identified for MCD, and it offers a key guidance for future studies seeking to integrate confidence estimation methods with gradient-based explainability methods. Namely, MCD may be incompatible with gradient-based explainability methods in some applications.

6. Recommendations for integration of confidence estimation approaches and explainability methods

While MCD has seen more widespread use than MCBN, its effects upon model gradients can prevent its effective integration with explainability methods, and it would be necessary to discard MCD iterations with NaNs in an implementation setting. As such, MCBN represents the more viable of the two methods for combination with gradient-based explainability methods. Within the context of applying MCBN with explainability methods, there are not large differences between explanations with MCBN relative to explanations for a regular model when those explanations are averaged across individuals and folds. As such, during model development, it may be unhelpful to integrate the two methods. However, we also found that on a per-individual basis, the spatial and temporal distributions of importance values associated with a traditional deep learning model tend to not be representative of the overall distribution that results when the model is combined with MCBN. As such, in a clinical implementation in which clinicians would be examining explanations for an individual patient and in which they would be evaluating model confidence estimates, a proper explanation of the confidence estimates would likely require an aggregate explanation that combined (e.g., via averaging) output explanations for each iteration of confidence estimation approaches. Additionally, some explainability methods, like saliency relative to LRP, tend to be noisier than others [52]. This tendency seems to be amplified when combined with confidence estimation approaches, as we found that generally the distributions of MCBN saliency were significantly different from their corresponding regular model importance values in more individuals than were those of LRP.

6.1. Limitations and next steps

While explainability is needed for the development of clinical decision support systems [11,37–39], a number of researchers have indicated that current explainability approaches are insufficient for use in a clinical setting [38,61]. There are valid concerns associated with their critiques. Nevertheless, for the purposes of this study, we only sought to provide a starting point to the integration of confidence estimation approaches and explainability methods. Further developments will be needed within the context of both confidence estimation and explainability in future years. However, they will eventually need to be integrated, and it is better that the field begin considering that integrative process sooner rather than later. While existing confidence estimation methods have had some popularity within the context of neuroimaging classification in recent years, they require repeated predictions and can be computationally intensive. Their repeated combination when directly paired with repeated output of explanations can be doubly intensive to the point of impracticality. This is, again, an example of how novel developments will be needed for both fields in the coming years.

7. Conclusion

In this study, we combined model confidence estimation approaches with explainability methods for the first time to help address the need for greater transparency in neuroimaging-based clinical decision support systems. We used two confidence estimation approaches –

MCD and MCBN. We further combined the two approaches with saliency and LRP for explainability. Our findings indicate that MCBN obtains comparable or better classification performance than MCD. Additionally, we found that MCD often adversely affected model gradients, while MCBN did not. We also uncovered spatial and temporal effects of SZ upon brain activity using our approach. It is our hope that this study will provide a starting point to the field on the integration of confidence estimation and explainability methods, provide useful guidance for future studies, and accelerate the development of transparent neuroimaging clinical decision support systems.

Acknowledgements

We thank those who collected the FBIRN dataset. Funding for this study was provided by NIH R01MH118695 and NSF 2112455.

References

[1]. Ellis CA, Miller RL, Calhoun VD. An approach for estimating explanation uncertainty in fMRI dFNC classification. bioRxiv; 2022.

[2]. Yan W, et al. Discriminating schizophrenia using recurrent neural network applied on time courses of multi-site FMRI data. EBioMedicine Sep. 2019;47:543–52. 10.1016/j.ebiom.2019.08.023. [PubMed: 31420302]

[3]. Rashid B, et al. Classification of schizophrenia and bipolar patients using static and dynamic resting-state fMRI brain connectivity. Neuroimage 2016;134:645–57. 10.1016/j.neuroimage.2016.04.051. [PubMed: 27118088]

[4]. Sen B, Mueller B, Klimes-Dougan B, Cullen K, Parhi KK. Classification of major depressive disorder from resting-state fMRI. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS, no. Mdd 2019:3511. 10.1109/EMBC.2019.8856453.-3514.

[5]. Chun JY, Sendi MSE, Sui J, Zhi D, Calhoun VD. Visualizing functional network connectivity difference between healthy control and major depressive disorder using an explainable machine-learning method. In: 2020 42nd annual international Conference of the IEEE Engineering in medicine & biology society. EMBC); 2020. p. 955–60. 10.1109/BIBE50027.2020.00162.

[6]. Challis E, Hurley P, Serra L, Bozzali M, Oliver S, Cercignani M. Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI. Neuroimage 2015. 10.1016/j.neuroimage.2015.02.037.

[7]. Liu S, Liu S, Cai W, Pujol S, Kikinis R, Feng D. Early diagnosis of ALZHEIMER'S disease with deep learning. 2014.

[8]. Ellis CA, Sattiraju A, Miller RL, Calhoun VD. Examining effects of schizophrenia on EEG with explainable deep learning models. bioRxiv; 2022. p. 5–8.

[9]. Ellis CA, Sattiraju A, Miller R, Calhoun V. Examining effects of schizophrenia on EEG with explainable deep learning models. bioRxiv; 2022. p. 5–8.

[10]. Gawne TJ, et al. A multimodal magnetoencephalography 7 T fMRI and 7 T proton MR spectroscopy study in first episode psychosis. Schizophr. Bull. 2020;6(1):1–9. 10.1038/s41537-020-00113-4.

[11]. Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inf Decis Making 2020;20(1):1–9. 10.1186/s12911-020-01332-6.

[12]. Bhatt U, et al. Uncertainty as a form of transparency: measuring, communicating, and using uncertainty," AIES 2021 - proc. AAAI/ACM Conf. AI, Ethics, Soc.; 2021. p. 401–13. 10.1145/3461702.3462571.2021.

[13]. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. 33rd Int. Conf. Mach. Learn. ICML 2016;3:1651–60. 2016.

[14]. Teye M, Azizpour H, Smith K. Bayesian uncertainty estimation for batch normalized deep networks. In: 35th international conference on machine learning, 11. ICML; 2018. p. 7824–33. 2018.

[15]. Rahman M, et al. Interpreting models interpreting brain dynamics. Sci Rep, 2022; 12:1–16. [PubMed: 34992227]

[16]. Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One Jul 2015;10(7). 10.1371/journal.pone.0130140.

[17]. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. Dec. 2013 [Online]. Available: http://arxiv.org/abs/1312.6034.

[18]. Zhang L. EEG signals classification using machine learning for the identification and diagnosis of schizophrenia. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS 2019:4521. 10.1109/EMBC.2019.8857946.– 4524.

[19]. Lebedev AV, et al. Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness. Neuroimage Clin. 2014. 10.1016/j.nicl.2014.08.023.

[20]. Böhle M, Eitel F, Weygandt M, Ritter K. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. Front Aging Neurosci 2019;11(Jul). 10.3389/fnagi.2019.00194.

[21]. Wood D, Cole J, Booth T. NEURO-DRAM: a 3D recurrent visual attention model for interpretable neuroimaging classification. Oct. 2019 [Online]. Available: http://arxiv.org/abs/1910.04721.

[22]. Sendi MSE, Chun JY, Calhoun VD. Visualizing functional network connectivity difference between middle adult and older subjects using an explainable machine- learning method. In: Proceedings - IEEE 20th international conference on bioinformatics and bioengineering, 2020. BIBE; 2020. p. 955–60. 10.1109/BIBE50027.2020.00162.

[23]. Zheng J, Wei X, Wang J, Lin H, Pan H, Shi Y. Diagnosis of schizophrenia based on deep learning using fMRI. Comput Math Methods Med 2021. 10.1155/2021/8437260.

[24]. Ellis CA, Sendi MSE, Miller RL, Calhoun VD. An unsupervised feature learning approach for elucidating hidden dynamics in rs-fMRI functional network connectivity. In: 2022 44th annual international conference of the IEEE engineering in medicine & biology society. EMBC); 2022. p. 4449–52.

[25]. Ellis CA, Sendi MSE, Geenjaar EPT, Plis SM, Miller RL, Calhoun VD. Algorithm- agnostic explainability for unsupervised clustering. 2021. p. 1–22 [Online]. Available: http://arxiv.org/abs/2105.08053.

[26]. Sendi MSE, et al. Aberrant dynamic functional connectivity of default mode network in schizophrenia and links to symptom severity. Front Neural Circ 2021; 15:1–14. 10.3389/fncir.2021.649417. March.

[27]. Liang M, et al. Widespread functional disconnectivity in schizophrenia with resting-state functional magnetic resonance imaging. Neuroreport 2006;17(2): 209–13. 10.1097/01.wnr.0000198434.06518.b8. [PubMed: 16407773]

[28]. Sendi MSE, Ellis CA, Milller RL, Salat DH, Calhoun VD. The relationship between dynamic functional network connectivity and spatial orientation in healthy young adults. bioRxiv; 2021.

[29]. Sendi MSE, et al. The link between brain functional network connectivity and genetic risk of Alzheimer's disease. bioRxiv 2021. 10.1002/alz.050101.

[30]. Ellis CA, Sancho ML, Miller R, Calhoun V. Exploring relationships between functional network connectivity and cognition with an explainable clustering approach. bioRxiv 2022:23–6.

[31]. Pedersen M, Verspoor K, Jenkinson M. Artificial intelligence for clinical decision support in neurology. Brain Commun. 2020;(–11):1. 10.1093/braincomms/fcaa096.

[32]. Williams LZJ, Fawaz A, Glasser MF, Edwards AD, Robinson EC. Geometric deep learning of the human connectome project multimodal cortical parcellation. In: Machine learning in clinical neuroimaging; 2021. p. 103–12. 10.1007/978-3-030-87586-2_11.

[33]. Shain C. CDRNN: discovering complex dynamics in human language processing. ACL-IJCNLP 2021 – 59th Annu. Meet. Assoc. Comput. Linguist. 11th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf. 2021:3718–34. 10.18653/v1/2021.acl-long.288.

[34]. Kia SM, Marquand AF. Neural processes mixed-effect models for deep normative modeling of clinical neuroimaging data. Proc. Mach. Learn. Res 2018:297–314 [Online]. Available: http://arxiv.org/abs/1812.04998.

[35]. Charitos AC, "Brain disease classification using multi-channel 3D convolutional neural networks," Linkoping University.

[36]. Yadav S, "Bayesian deep learning based convolutional neural network for classification of Parkinson's disease using functional magnetic resonance images".

[37]. Vieira S, Pinaya WHL, Mechelli A. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications. Neurosci Biobehav Rev Mar. 01 2017;74:58–75. 10.1016/j.neubiorev.2017.01.002. Elsevier Ltd. [PubMed: 28087243]

[38]. Nazar M, Alam MM, Yafi E, Su'Ud MM. A systematic Review of human-computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques. IEEE Access 2021;9:153316–48. 10.1109/ACCESS.2021.3127881.

[39]. Gerlings J, Jensen MS, Shollo A, Explainable AI. But explainable to whom? An exploratory case study of xAI in healthcare. Intell. Syst. Ref. Libr 2022;212: 169–98. 10.1007/978-3-030-83620-7_7. Ml.

[40]. Thomas AW, Heekeren HR, Müller K-R, Samek W. Analyzing neuroimaging data through recurrent deep learning models. Front. Neurosci; Oct. 2019 [Online]. Available: http://arxiv.org/abs/1810.09945.

[41]. Qiao K, et al. Accurate reconstruction of image stimuli from human functional magnetic resonance imaging based on the decoding model with capsule network architecture. Front Neuroinf Sep 2018;12. 10.3389/fninf.2018.00062.

[42]. Yan Y, Solarz E, Zhu J, Sripada C, Duda M, Koutra D. Groupinn: grouping-based interpretable neural network for classification of limited, noisy brain data. Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min 2019:772–82. 10.1145/3292500.3330921. July.

[43]. Li X, et al. BrainGNN: interpretable brain graph neural network for fMRI analysis. Med Image Anal 2021;74:102233. 10.1016/j.media.2021.102233.

[44]. Jiang Z. et al. , "Attention module improves both performance and interpretability of 4D fMRI decoding neural network," arXiv, no. Dl.

[45]. Du Y, et al. NeuroMark: an automated and adaptive ICA based pipeline to identify reproducible fMRI markers of brain disorders. Neuroimage Clin. 2020;28(August): 102375. 10.1016/j.nicl.2020.102375.

[46]. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: 3rd International Conference on Learning Representations (ICLR); 2015.

[47]. Lin QH, Niu YW, Sui J, Da Zhao W, Zhuo C, Calhoun VD. SSPNet: an interpretable 3D-CNN for classification of schizophrenia using phase maps of resting-state complex-valued fMRI data. Med Image Anal 2022;79:102430. 10.1016/j.media.2022.102430.

[48]. Vilamala A, Madsen KH, Hansen LK. Deep convolutional neural networks for interpretable analysis of EEG sleep stage scoring. IEEE Int. Work. Mach. Learn. Signal Process. MLSP 2017:1–6. 10.1109/MLSP.2017.8168133. 2017-Septe, no. 659860.

[49]. Frick T, Glüge S, Rahimi A, Benini L, Brunschwiler T. Explainable deep learning for medical time series data. Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST 2021;362: 244–56. LNICST.

[50]. Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One Jul 2015;10(7). 10.1371/journal.pone.0130140.

[51]. Ancona M, Ceolini E, Oztireli C, Gross M. Towards better understanding of ¨ gradient-based attribution methods for deep neural networks. In: International conference on learning representations; 2018. p. 1–16.

[52]. Samek W, Binder A, Montavon G, Lapuschkin S, Müller KR. Evaluating the visualization of what a deep neural network has learned. IEEE Transact Neural Networks Learn Syst 2017;28(11):2660–2673, Nov. 10.1109/TNNLS.2016.2599820.

[53]. Yan W, et al. Discriminating schizophrenia from normal controls using resting state functional network connectivity: a deep neural network and layer-wise relevance propagation method. 2017.

[54]. Ellis CA, Miller RL, Calhoun VD. A systematic approach for explaining time and frequency features extracted by CNNs from raw EEG data. 2022. bioRxiv.

[55]. Ellis CA, et al. Novel methods for elucidating modality importance in multimodal electrophysiology classifiers. bioRxiv; 2022.

[56]. Ellis CA, Sendi MS, Willie JT, Mahmoudi B. Hierarchical neural network with layer-wise relevance propagation for interpretable multiclass neural state classification. In: 10th international IEEE/EMBS conference on neural engineering. NER); 2021. p. 18–21.

[57]. Lemay A, et al. arXiv 2021:1–6 [Online]. Available: http://arxiv.org/abs/2111.06754.

[58]. Sendi MSE, et al. Multiple overlapping dynamic patterns of the visual sensory network in schizophrenia. Schizophr Res 2021;228:103–11. 10.1016/j.schres.2020.11.055. [PubMed: 33434723]

[59]. Yamamoto M, et al. Aberrant functional connectivity between the thalamus and visual cortex is related to attentional impairment in schizophrenia. Psychiatry Res Neuroimaging 2018;278:35–41. 10.1016/j.pscychresns.2018.06.007. June. [PubMed: 29981940]

[60]. Chen X, et al. Functional disconnection between the visual cortex and the sensorimotor cortex suggests a potential mechanism for self-disorder in schizophrenia. Schizophr Res 2014;166(1–3):151–7. 10.1016/j.schres.2015.06.014.

[61]. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit. Heal 2021;3(11): e745. 10.1016/S2589-7500(21)00208-9. –e750.

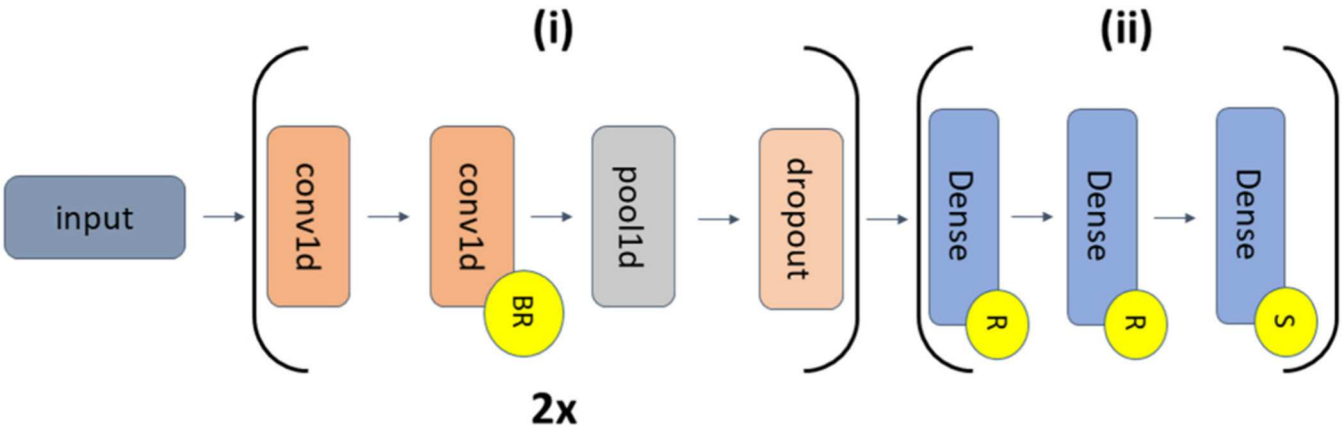

Fig. 1.

CNN Architecture. The model has sections (i) and (ii) for feature extraction and classification, respectively. Layers in section (i) are repeated twice for different hyperparameters. The first and second pairs of convolutional layers (conv1d) have a kernel size of 10 and 16 and 24 filters, respectively. Each pair of conv1d layers is followed by a max pooling layer with a pool size of 2 and spatial dropout (rates = 0.3 and 0.4). Layers in section (ii) include 3 dense layers with 10, 6, and 2 nodes. Yellow circles with "BR", "R", and "S" correspond to layers with batch normalization/ReLU, ReLU, and softmax activations, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

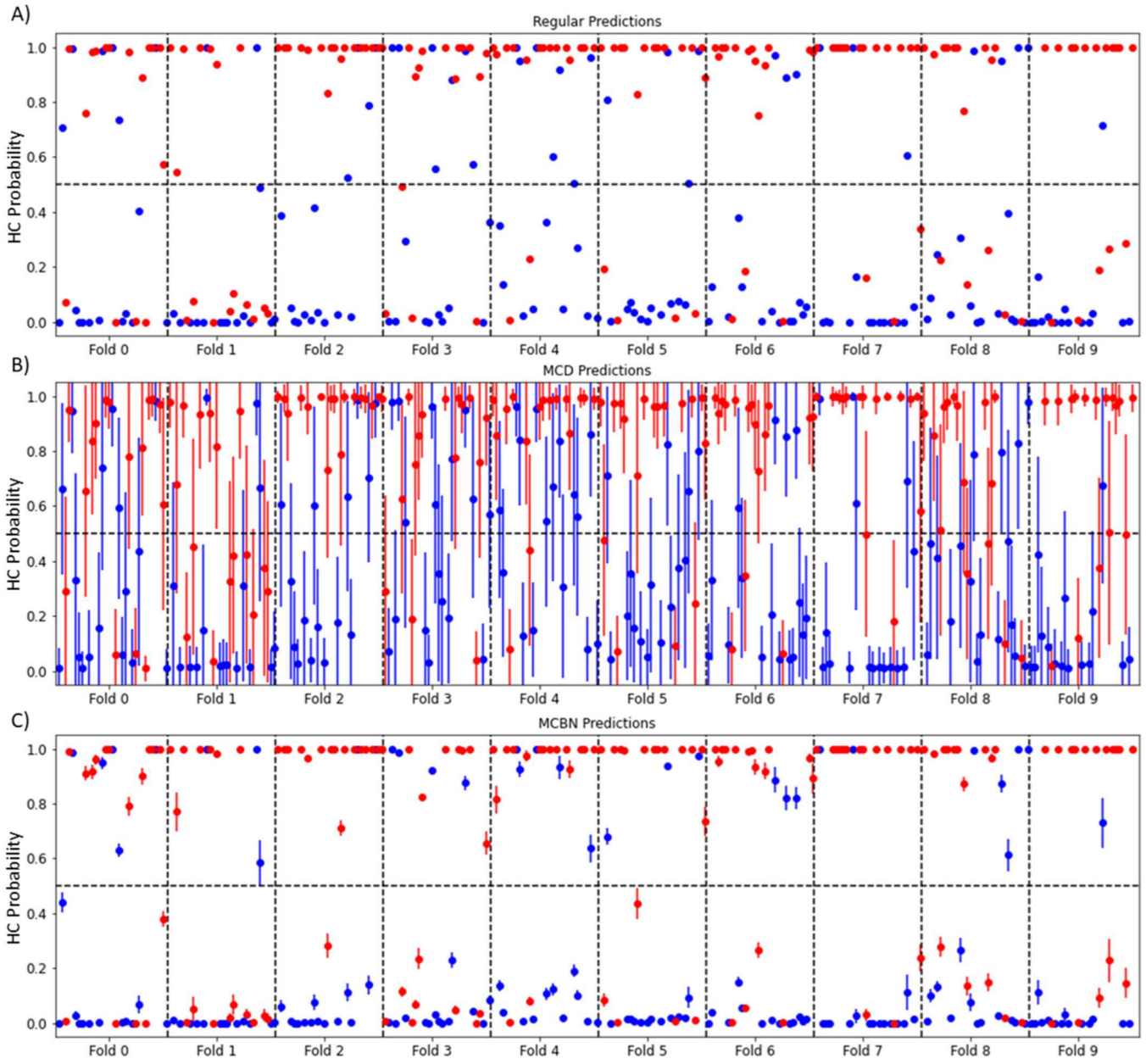

Fig. 2.

Distributions of Sample Predictions. Panel A shows the model predictions without confidence estimation. Panels B) and C) show the model predictions with MCD and MCBN, respectively. The predictions for samples with true labels of HC and SZ are shown in red and blue respectively, and samples are aligned in the same order across panels such that each panel can be visually compared. It should be noted that the points in Panels B and C reflect the mean of predictions, and the error lines reflect one standard deviation above and below the mean. Samples are grouped from left to right based on their folds, with a black dashed vertical line separating samples for each fold. The y-axis reflects the probability of a sample belonging to the HC class, and the black dashed horizontal line indicates the 50% boundary point between classes. As such, blue samples above the boundary point are misclassified,

and red samples below the boundary point are misclassified. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

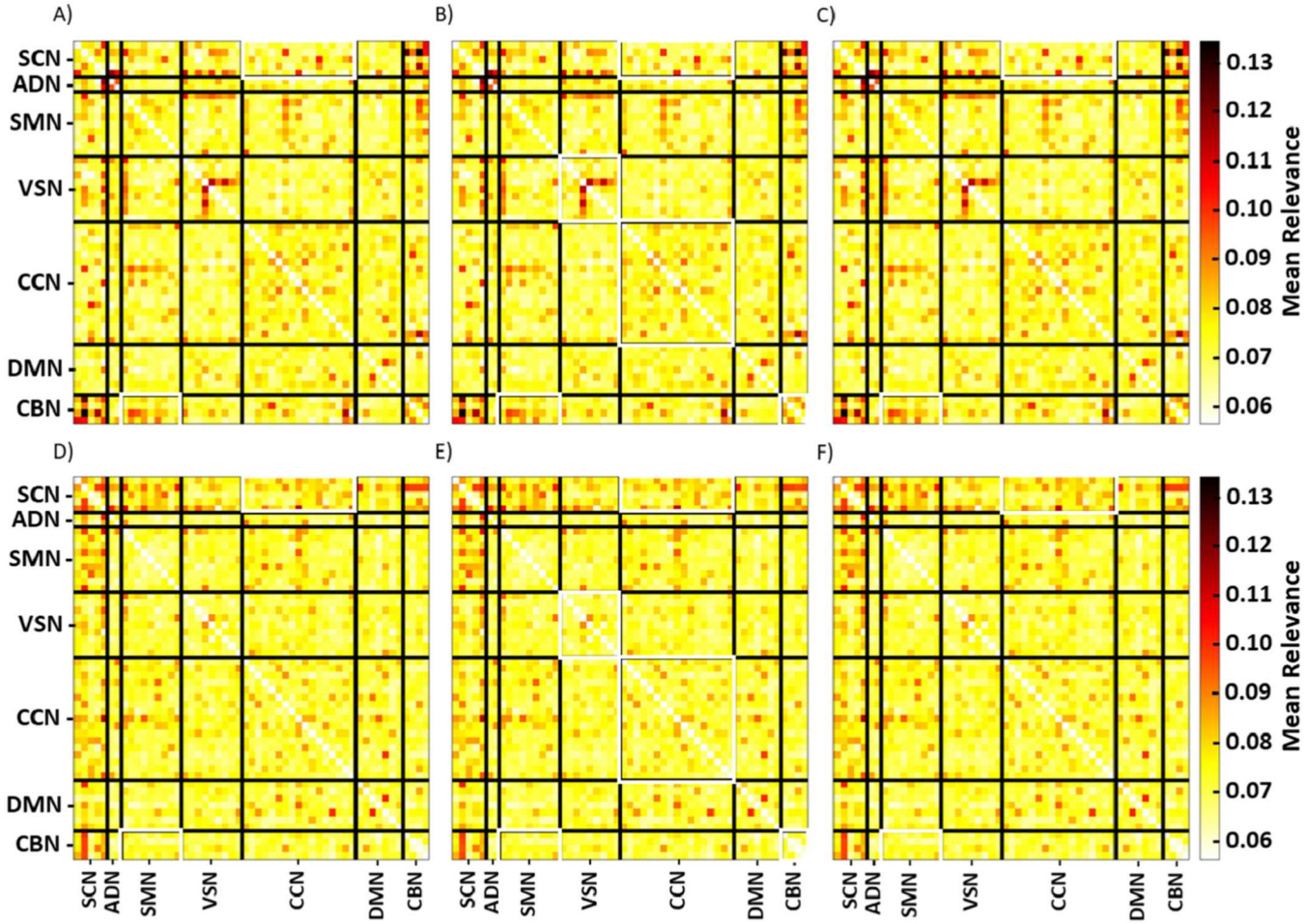

Fig. 3.

Mean of Total LRP Relevance Across All Timesteps. Panels A, B, and C reflect the mean relevance of the regular model, the model with MCD, and the model with MCBN for SZs. Panels D, E, and F show the same values for HCs. Networks are included on the x- and y-axes and are separated by black lines. Network pairs surrounded by white boxes have statistically significant differences between HCs and SZs. Lastly, all panels share the color bars to the right of Panels C and F. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

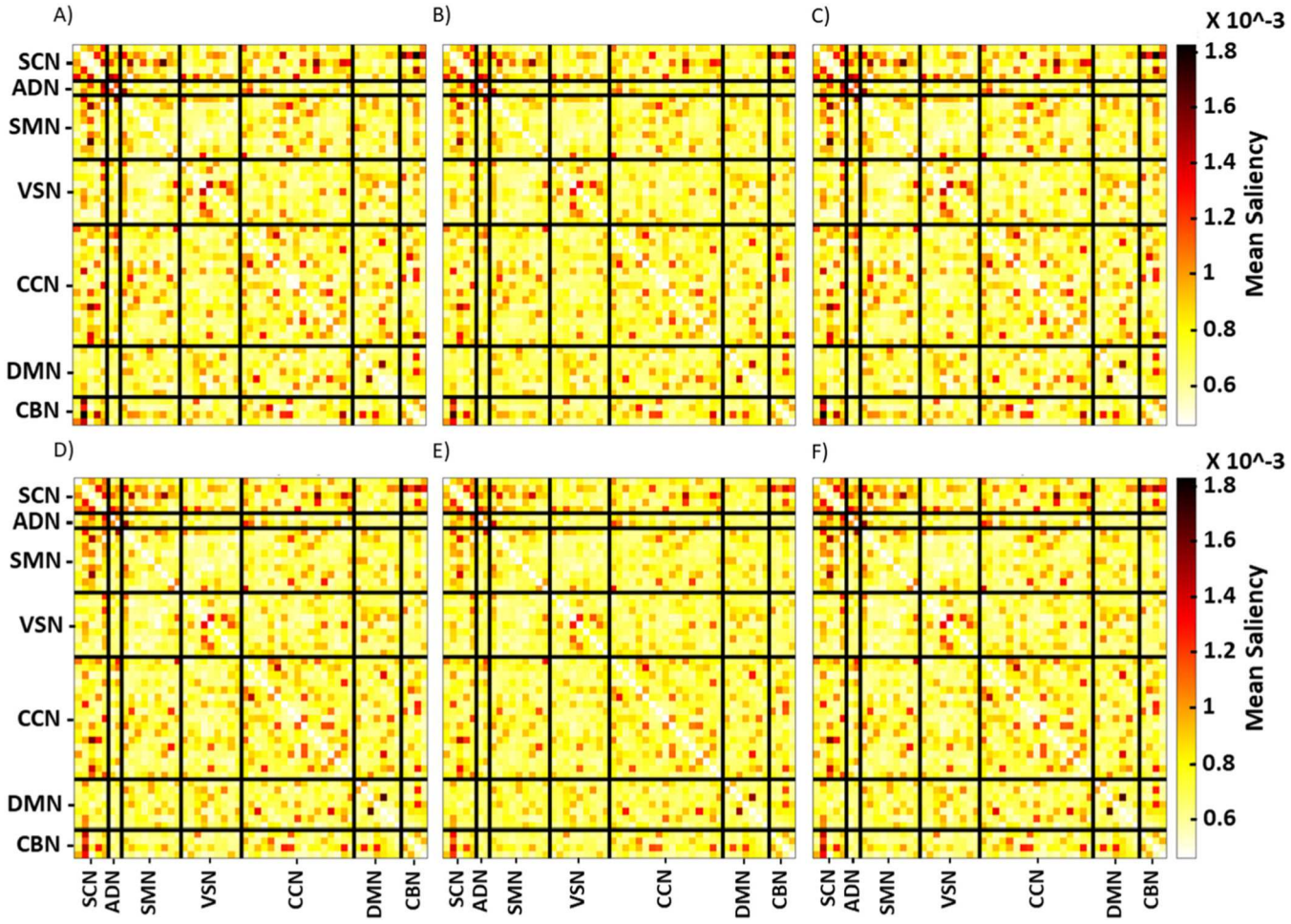

Fig. 4.

Mean of Total Saliency Across All Timesteps. Panels A, B, and C reflect the mean saliency of the regular model, the model with MCD, and the model with MCBN for SZs. Panels D, E, and F show the same values for HCs. Networks are included on the x- and y-axes and are separated by black lines. All panels share the color bars to the right of Panels C and F. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

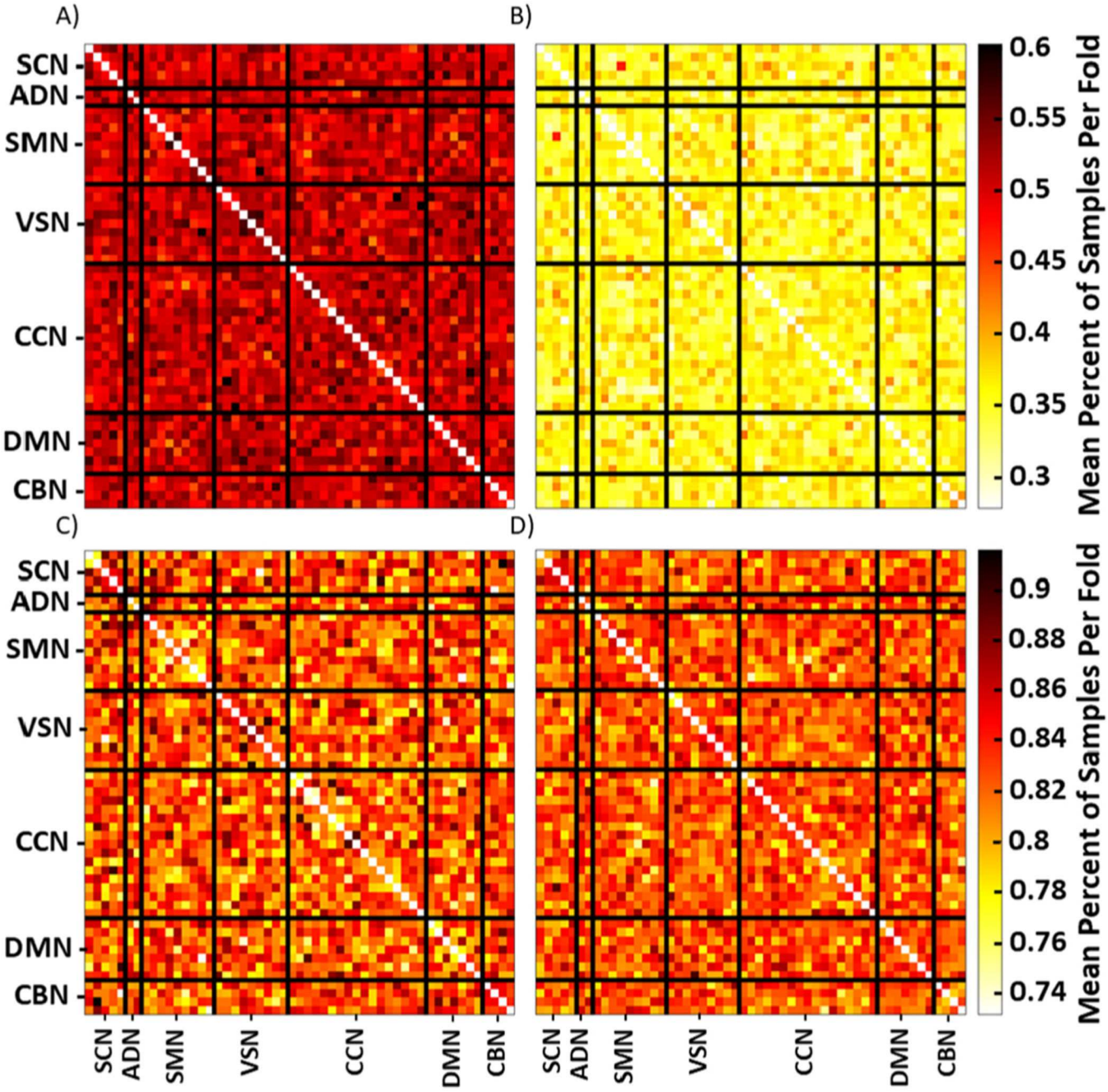

Fig. 5.

Sample Level Differences in Spatial Importance. Panels A and B show the mean percent of samples per fold with differences ($p < 0.05$) between their regular relevance values and their MCD and MCBN relevance distributions, respectively. Panels C and D show the mean percent of samples per fold with differences ($p < 0.05$) between their regular saliency values and their MCD and MCBN saliency distributions, respectively. The color bars to the right of Panels B and D are shared by Panels A and B and Panels C and D, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

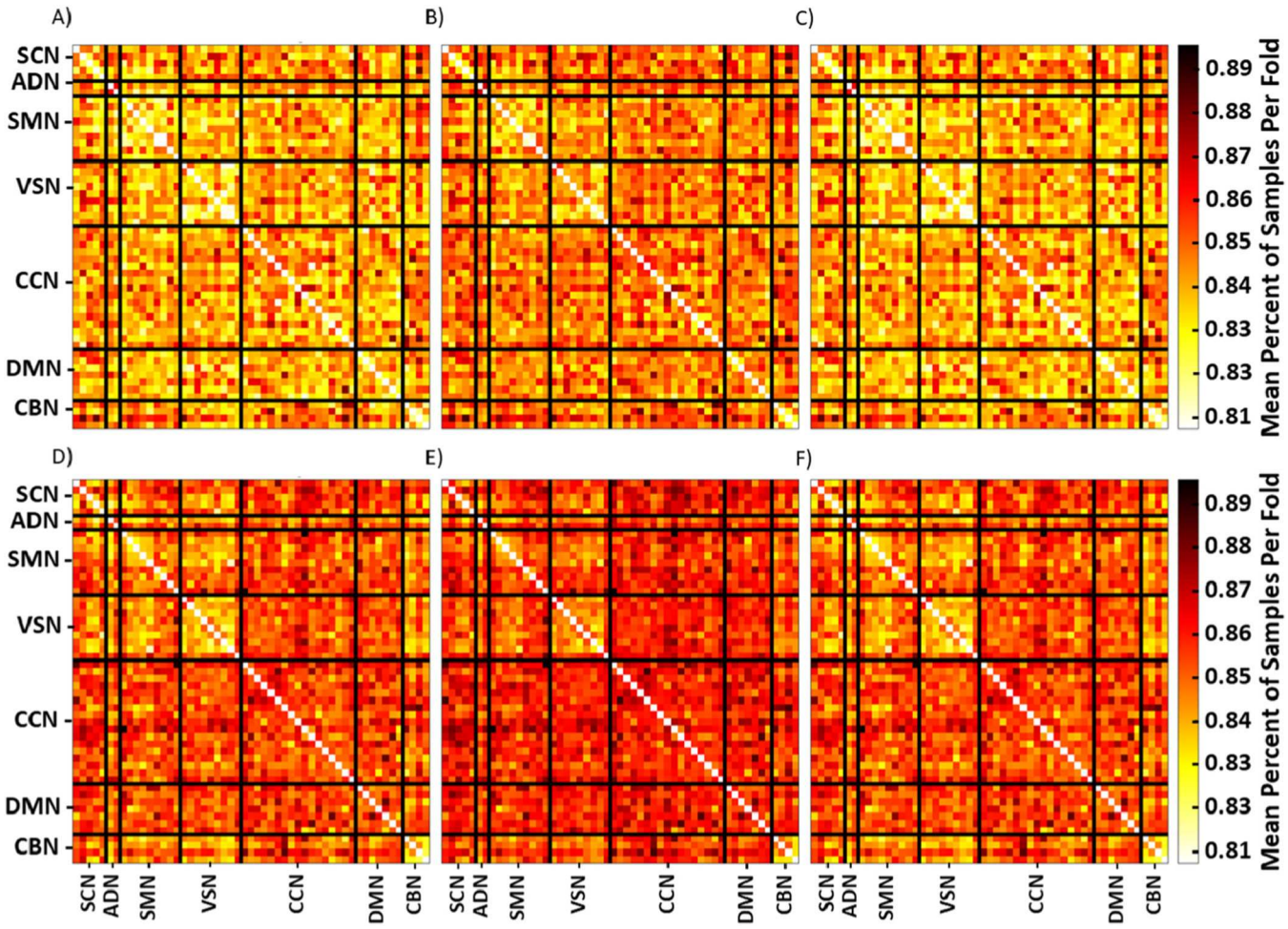

Fig. 6.

Mean of Relevance EMD over Time. Panels A, B, and C reflect the mean EMD of the regular model, the model with MCD, and the model with MCBN for SZs. Panels D, E, and F show the same values for HCs. Networks are included on the x- and y-axes and are separated by black lines. All panels share the color bars to the right of Panels C and F. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

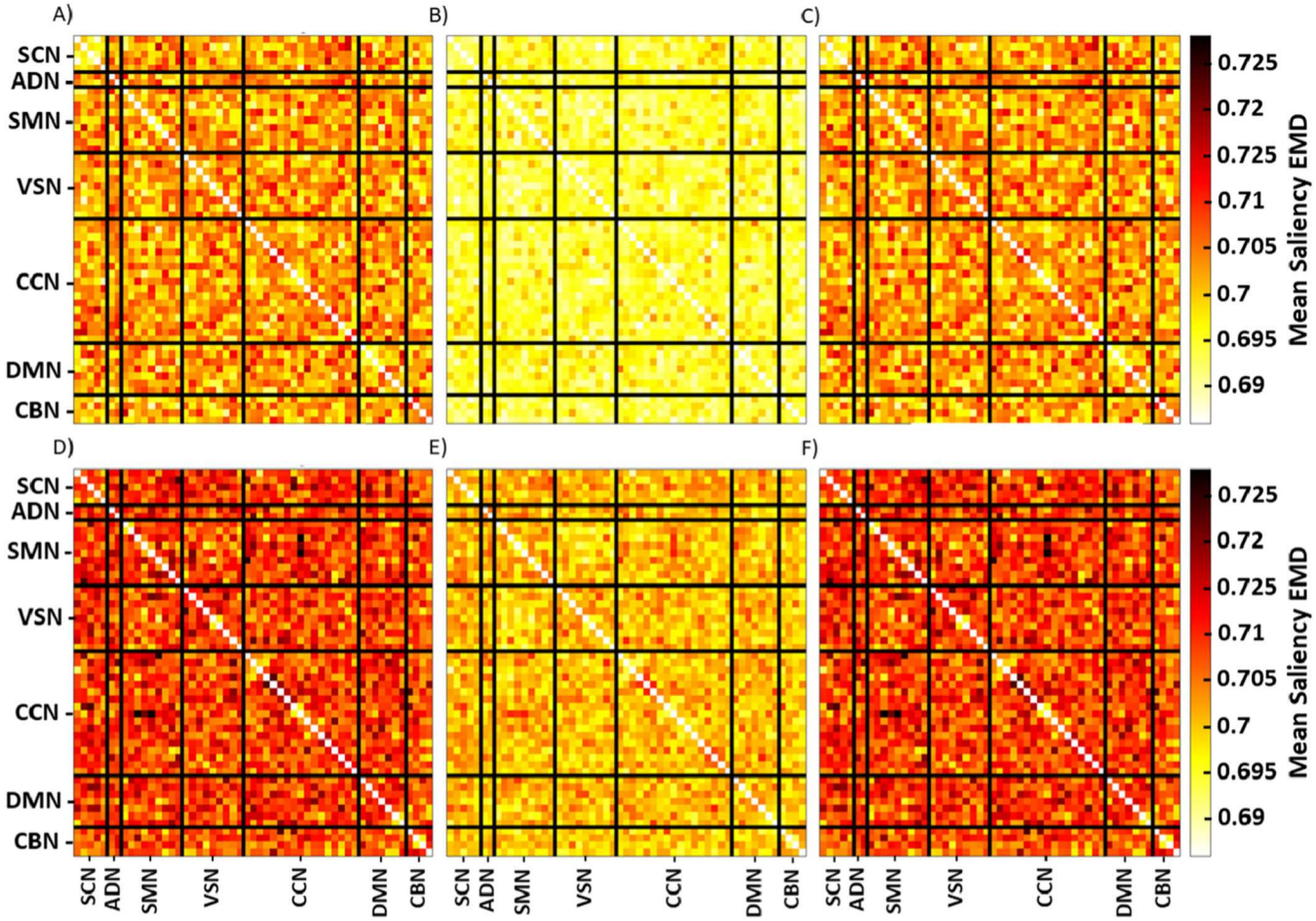

Fig. 7.

Mean of Saliency EMD over Time. Panels A, B, and C reflect the mean EMD of the regular model, the model with MCD, and the model with MCBN for SZs. Panels D, E, and F show the same values for HCs. Networks are included on the x- and y-axes and are separated by black lines. All panels share the color bars to the right of Panels C and F. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

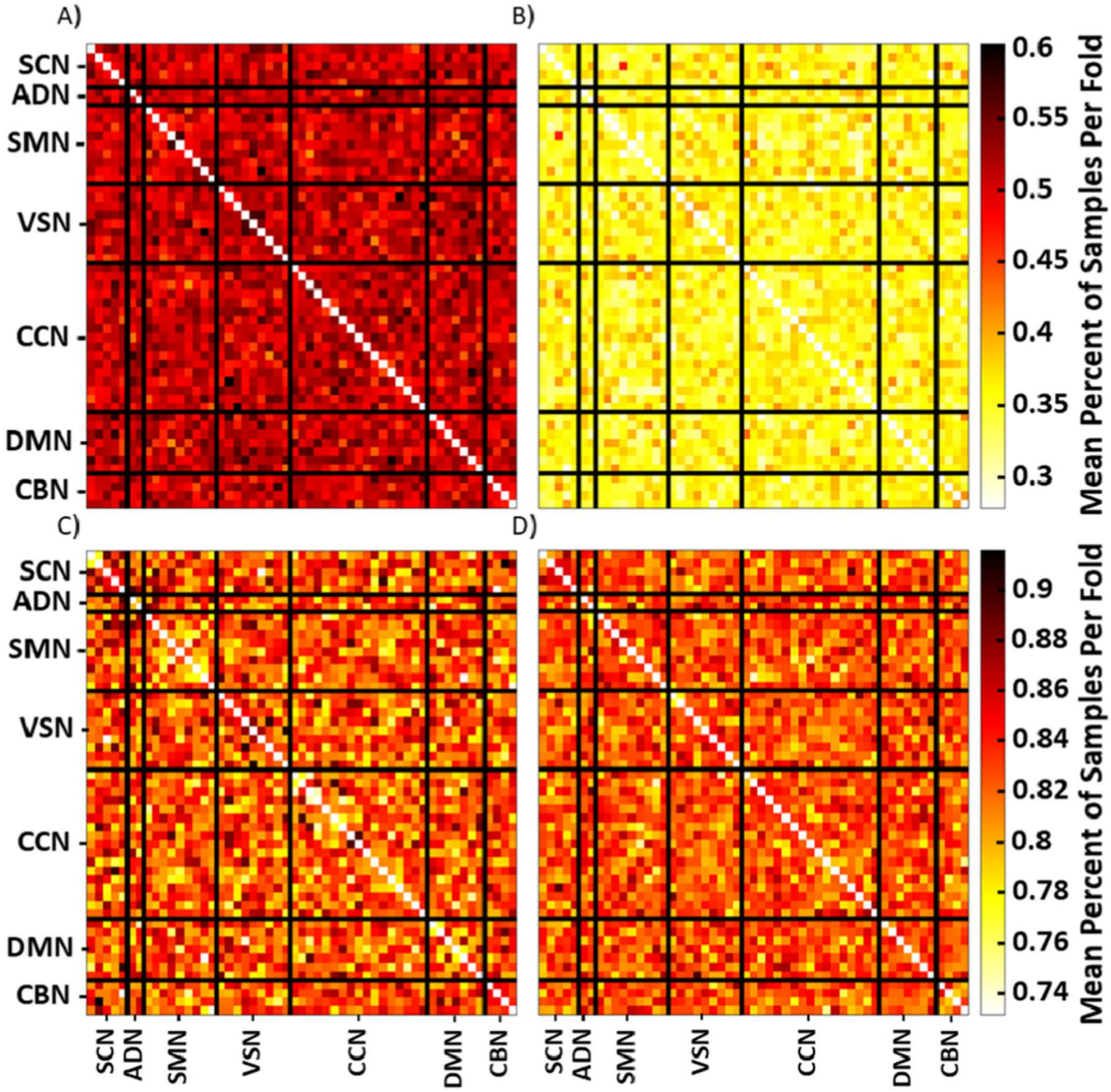

Fig. 8.

Sample Level Differences in Temporal Importance Distributions. Panels A and B show the mean percent of samples per fold with differences ($p < 0.05$) between their regular relevance EMD values and their MCD and MCBN EMD distributions, respectively. Panels C and D show the mean percent of samples per fold with differences ($p < 0.05$) between their regular saliency EMD values and their MCD and MCBN saliency EMD distributions, respectively. The color bars to the right of Panels B and D are shared by Panels A and B and Panels C and D, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

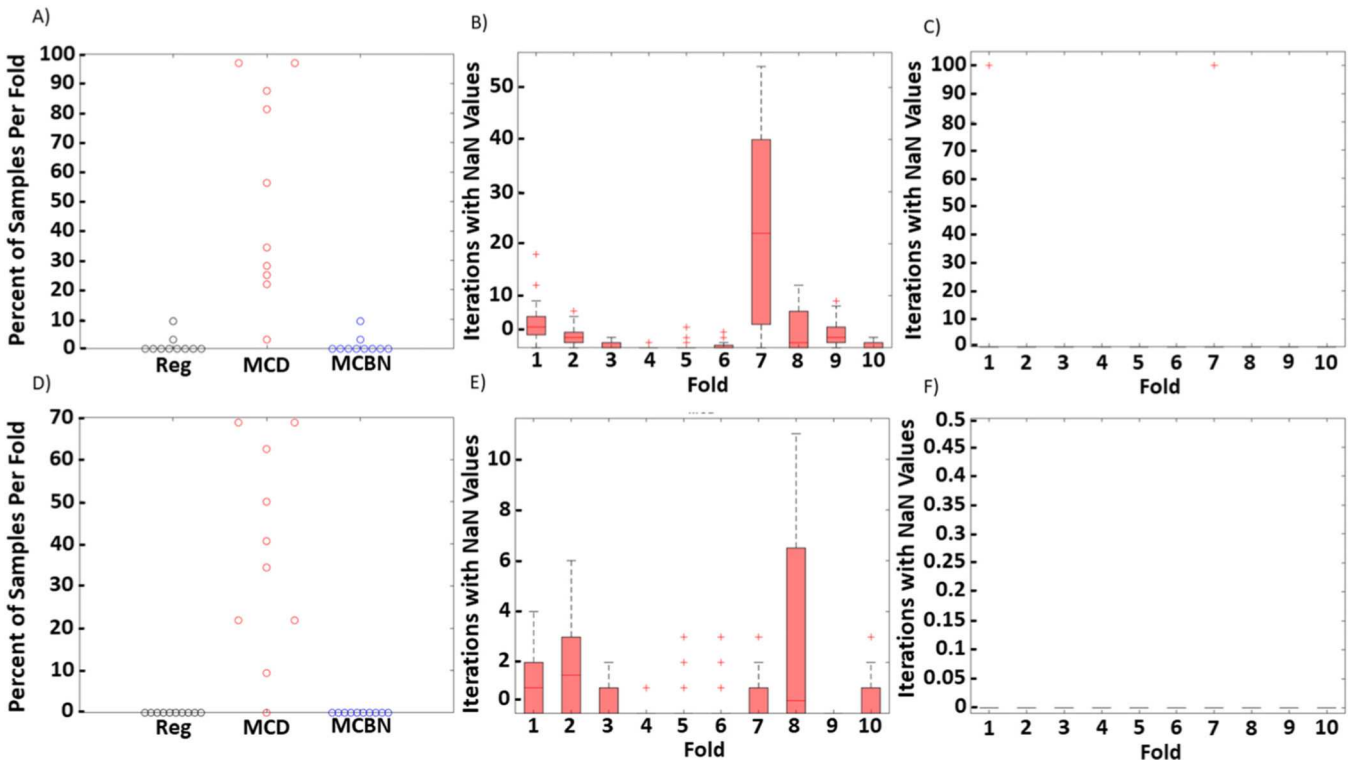

Fig. 9.

Distribution of NaN Values. Panels A and D show the percent of samples per fold with at least one NaN value for LRP and saliency, respectively. The values for the regular model, the model with MCD, and the model with MCBN are shown in black, red, and blue, respectively. Panels B and C show the percent of iterations per sample of each fold that produced NaN LRP relevance values for MCD and MCBN, respectively. Panels E and F show the percent of iterations per sample of each fold that produced NaN saliency values for MCD and MCBN, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 1

Model performance results.

	SPEC	SENS	ACC
Regular	75.63 ± 14.64	74.38 ± 11.34	75.00 ± 07.26
MCBN	70.63 ± 15.32	79.35 ± 06.87	75.00 ± 09.06
MCD	78.13 ± 13.76	68.13 ± 13.76	73.13 ± 07.93