NeuroImage 241 (2021) 118402

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

MEGnet: Automatic ICA-based artifact removal for MEG using spatiotemporal convolutional neural networks

Alex H. Treacher[a,1], Prabhat Garg[a,b,1], Elizabeth Davenport[b,d], Ryan Godwin[e], Amy Proskovec[b,d], Leonardo Guimaraes Bezerra[e], Gowtham Murugesan[b], Ben Wagner[b,d], Christopher T. Whitlow[e], Joel D. Stitzel[e], Joseph A. Maldjian[a,d], Albert A. Montillo[a,b,c,*]

[a] Lyda Hill Department of Bioinformatics, UT Southwestern Medical Center, Dallas, TX, United States
[b] Department of Radiology, UT Southwestern Medical Center, Dallas, TX, United States
[c] Advanced Imaging Research Center, UT Southwestern Medical Center, Dallas, TX, United States
[d] Magnetoencephalography Center of Excellence, UT Southwestern Medical Center, Dallas, TX, United States
[e] Wake Forest School of Medicine, Winston-Salem, NC, United States

ARTICLE INFO

Keywords:
MEG
Artifact
Automation
ICA
Convolutional neural network
Deep learning

ABSTRACT

Magnetoencephalography (MEG) is a functional neuroimaging tool that records the magnetic fields induced by neuronal activity; however, signal from non-neuronal sources can corrupt the data. Eye-blinks, saccades, and cardiac activity are three of the most common sources of non-neuronal artifacts. They can be measured by affixing eye proximal electrodes, as in electrooculography (EOG), and chest electrodes, as in electrocardiography (ECG), however this complicates imaging setup, decreases patient comfort, and can induce further artifacts from movement. This work proposes an EOG- and ECG-free approach to identify eye-blinks, saccades, and cardiac activity signals for automated artifact suppression.

The contribution of this work is three-fold. First, using a data driven, multivariate decomposition approach based on Independent Component Analysis (ICA), a highly accurate artifact classifier is constructed as an amalgam of deep 1-D and 2-D Convolutional Neural Networks (CNNs) to automate the identification and removal of ubiquitous whole brain artifacts including eye-blink, saccade, and cardiac artifacts. The specific architecture of this network is optimized through an unbiased, computer-based hyperparameter random search. Second, visualization methods are applied to the learned abstraction to reveal what features the model uses and to bolster user confidence in the model's training and potential for generalization. Finally, the model is trained and tested on both resting-state and task MEG data from 217 subjects, and achieves a new state-of-the-art in artifact detection accuracy of 98.95% including 96.74% sensitivity and 99.34% specificity on the held out test-set. This work automates MEG processing for both clinical and research use, adapts to the acquired acquisition time, and can obviate the need for EOG or ECG electrodes for artifact detection.

1. Introduction

Magnetoencephalography (MEG), is a functional neuroimaging method that offers better temporal resolution than fMRI (Bellec et al., 2010; Dekhil et al., 2018; Duan et al., 2013; Fatima et al., 2013). MEG also uses a more direct measure of neuronal activity via the magnetic flux induced by neuronal activity compared to fMRI, which measures activity indirectly through the blood-oxygen-level-dependent (BOLD) response that can be compromised through vascular decoupling. Compared to electroencephalography (EEG) which is also a direct measure of neuronal activity, MEG's reliance upon magnetic flux rather than electrical conduction is advantageous as the flux is less affected by intervening tissue characteristics and can yield more accurate source space reconstruction (Buzsáki et al., 2012; Fatima et al., 2013; Muthukumaraswamy, 2013). Nevertheless, MEG is vulnerable to noise from non-neuronal sources. For example, the spectral bandwidth of muscle activity overlaps with the gamma-frequency band of neuronal activity (Criswell and Cram, 2011; Muthukumaraswamy, 2013). In particular, eye-blink (EB) artifacts, saccade (SA) artifacts, and cardiac activity (CA) artifacts, which are three of the most common sources of artifact in MEG data, share frequency bands (1 Hz – 20 Hz) with alpha, theta, and delta brain waves (Breuer et al., 2014; Zikov et al., 2002). Fig. 1 shows how such artifacts can corrupt much of brain source space when recon-

* Corresponding author.
E-mail address: Albert.Montillo@UTSouthwestern.edu (A.A. Montillo).
[1] These authors contributed equally.

https://doi.org/10.1016/j.neuroimage.2021.118402.
Received 1 December 2020; Received in revised form 18 June 2021; Accepted 15 July 2021
Available online 16 July 2021.

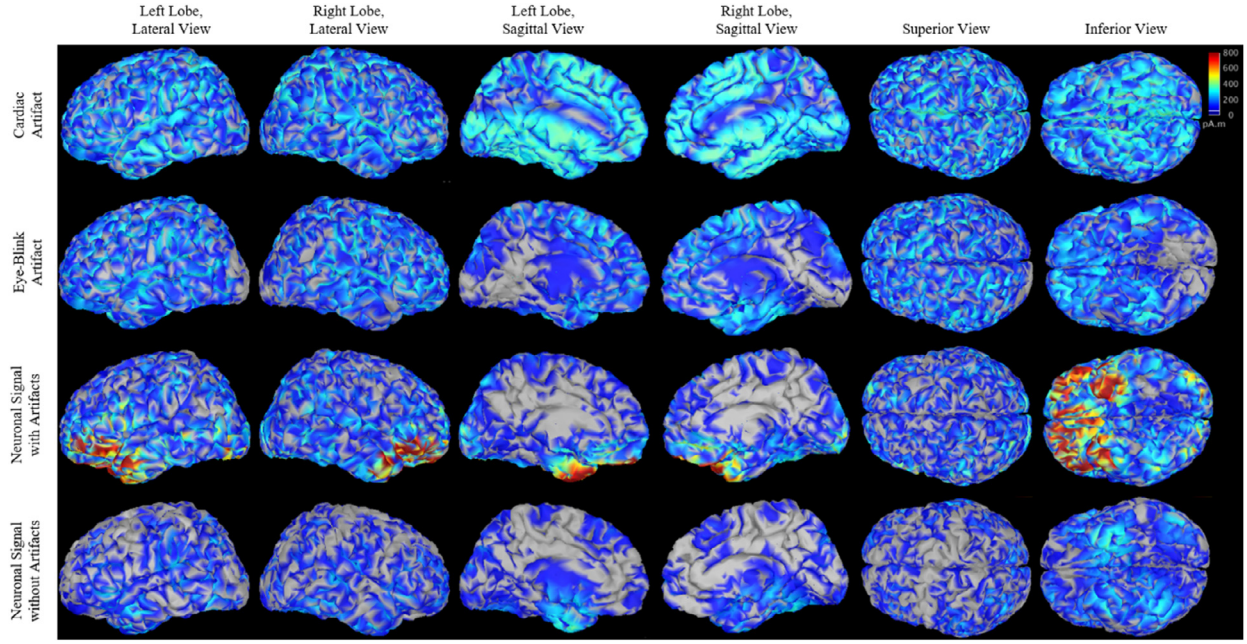

Fig. 1. The manifestation of MEG artifacts in brain space, reconstructed using minimum norm estimate (MNE) source localization. First row: projection of isolated cardiac artifact ICA component. Second row: projection of isolated eye-blink artifact component. Third row: projection of all 20 ICA components, including artifacts. Fourth row: projection of only the all neuronal ICA components without the cardiac and eye-blink components. Both cardiac and eye-blink artifact projections can demonstrate diffuse activity across much of brain space. The amplitudes in third row are much higher than fourth row due to the effects of artifacts on source space.

structed directly from the MEG sensor signals, making artifact identification and suppression crucial for mapping true brain activity.

Manually removing artifacts from MEG using Independent Component Analysis (ICA) can improve MEG signal-to-noise ratio by up to 35% on task MEG and it has been suggested that these conclusion hold for resting-state MEG (Gonzalez-Moreno et al., 2014). ICA is a source separation method that decomposes the data into individual independent components, separating artifact and signal in the process. However, these components are randomly ordered and must be manually labeled as neuronal signal or artifact (Gross et al., 2013; Muthukumaraswamy, 2013) allowing the neuronal components to be projected back into sensor space. Manual labeling of artifacts for MEG processing is prohibitive as it is both time consuming and requires a MEG expert. In addition, manual labeling is subjective as it is dependent on the rater's experience, which can decrease the reproducibility of MEG processing. To automate the detection of EB and CA artifacts, some researchers use electrooculography (EOG) and electrocardiography (ECG) electrodes to separately record the eye-blink and cardiac artifact signals (Breuer et al., 2014). However, these methods can add complexity and time to the data acquisition setup (especially important for large cohorts), can be uncomfortable for the some subjects such as those with sensitive skin, and may induce additional artifacts from postural muscle movements and facial twitching. Additionally, although ICA components can be ranked based on correlation with the signal from the EOG and ECG electrodes, manual labeling is still required in commonly used pipelines (Tutorials/Epilepsy - Brainstorm, 2021). This work presents an automated, objective, pipeline that can detect EB, CA and SA artifacts in MEG data that does not require EOG- and ECG-recordings.

This work aims to automate the removal of ICA components to increase the signal to noise ratio of MEG data, and make MEG data more readily useable. To achieve this we build a highly accurate, generalizable, unbiased, and adaptable model to automate the detection and removal of ICA artifacts in MEG. In addition, the general framework is made open source such that others can create a custom pipeline for their work, if needed. In our previous work, a neural network was built to detect eye-blink artifacts using the ICA derived spatial maps (Garg et al., 2017b). Later, models were designed that can detect eye-blink and cardiac artifacts using the ICA derived time courses (Garg et al., 2017a). Others have also published work automating the detection of artifactual ICA components, these works are compared to this work in Section 4.1. This research builds on our past work, and provides various improvements to prior work published by others in several important aspects including; an increase in performance, an increase in model generalizability and reliability, and generation of a highly optimized, custom model. To achieve these objectives (1) a large dataset is formed for model training consisting of both resting-state and task-based MEG. This training dataset includes subjects from 3 different databases, span a large age range of 15 to 73 year old subjects, and includes both sexes. (2) The model is trained to seamlessly integrate both ICA spatial maps and the time courses for artifact detection. (3) Models are built to that take advantage of the available acquired MEG data regardless of the acquisition duration: e.g. 1–80 min. (4) Held out test performance and validation performance, rather than just validation performance is reported to facilitate results comparison. (5) The model is optimized using an extensive automated neural architecture search. (6) Ground truth is formed from the consensus of 4 expert raters. (7) Finally, the model is shown to be highly interpretable and understandable through an analysis revealing what parts of the spatial maps and time courses are used for artifact detection.

2. Materials and methods

2.1. Magnetoencephalography data

This study uses both resting-state and task-based MEG data from 294 scans from 217 subjects with ages ranging from 10 years to 73 years. To include both resting-state and task MEG, both sexes, and achieve such a wide age distribution, data is drawn from three databases described in the following sections. Demographics for the subjects in the overall MEG training dataset are summarized in Table 1.

From the 217 subject dataset, 46 subjects and a total of 62 scans (20% of the data) are set aside prior to training for testing the model's performance, while the remaining 80% or 171 subjects with a total of

Table 1
Demographics of each database, and the combined dataset used in this work.

Database	McGill	iTAKL	HCP	Combined
Subject Count	82	49	86	217
Age: Mean(std)	34.6(9.33)	13.3(2.63)	28.8(3.24)	28.5(8.87)
Age: min/max	28/73	10/18	23.5/33	10/73
Sex: M/F	41/46	49/0	74/11	164/57

232 scans are subsequently partitioned by 10-fold cross validation for model training and hyperparameter optimization. The winning model is selected and its performance is evaluated on the held out test set not used during training or model selection. All splits including the test split, and the folds of the 10-fold cross validation are similarly stratified by age, sex, site (origin database), and scan type (resting-state or task-based), and grouped by subject. Statistical tests, including a Student's T-test is for continuous-valued age, and the Chi square test used for sex, site, and scan type, reveal the success of the stratified partitioning with no statistically significant differences ($p < .05$) between splits (Supplemental Table S1).

2.1.1. Database 1: McGill OMEGA dataset

Five minutes of continuous resting-state, eyes-open MEG data was obtained from 82 volunteers within the Open MEG Archive (OMEGA) (Nisoet al., 2016). This dataset contains subjects with ages 18 through 73 years and is 53% female. The participants were instructed to look at a target (fixate) during the acquisition. The McGill OMEGA dataset was collected on CTF whole-head MEG system (VSM MedTech Ltd., Coquitlam, Canada) that consist of 275 first-order axial-gradiometer coils (ctfmeg, 2020.000Z). MEG signals were sampled at a rate of 2400 Hz and a bandwidth of 1–80 Hz.

2.1.2. Database 2: Imaging telemetry and kinematic modeling in youth football (iTAKL) dataset

Eight minutes of continuous resting-state MEG data was obtained from 49 male football players: 30 youth (10–13 years) and 19 high school (14–18 years old) subjects, as part of the Imaging Telemetry And Kinematic modeling in youth football (iTAKL) concussion study (Davenport et al., 2014). The participants were instructed to look at a target (fixate) during the acquisition. MEG signals were recorded using a 275 channel axial gradiometer whole-head CTF system with 29 additional reference sensors for noise cancelation, and sampled at a rate of 600 Hz with an acquisition bandwidth of 0.25–150.

2.1.3. Database 3: Human connectome project (HCP) dataset

Task data was obtained from the HCP database (van Essen et al., 2012). Scan times varied between 7 and 13 min. Three different tasks are available and all 3 were used in this study: a sensory motor task, a working memory task, and a language processing (story memory) task. To provide a roughly balanced number of resting-state and task-based MEG scans for training, a total of 150 task-based scans were selected from 89 different subjects. The 150 scans were selected by maximizing the number of subjects, while also maximizing age range, and balancing the sex and number of scans per task. HCP used a MAGNES 3600 MEG system with 248 magnetometer channels, 23 reference channels, a sapling rate of 2034.5101 Hz, and a bandwidth of 1–90 Hz. Additional information on the task data is available in (Larson-Prior et al., 2013).

2.2. Overview of the proposed MEG pipeline

An overview of the proposed MEG processing pipeline is shown in Fig. 2. Constituent electrical activity in the physical space of the MEG scanner (Fig. 2a) include eye-blinks, saccades, and cardiac activity as well as true neuronal activity. These activities induce magnetic flux measured by the MEG sensors near the scalp. Raw MEG sensor space data is corrupted by the non-neuronal activity and can manifest as large troughs and perturbations in the sensor space data (Fig. 2b, left, red arrows). Consequently, a naïve reconstruction of brain space activity, where raw data is projected into source space without prior artifact removal, does not estimate well the true neuronal activity (Fig. 2b, right). The proposed pipeline applies preprocessing, and extracts independent components (Fig. 2c) via ICA in steps whose details are described in the following sections. The proposed MEGnet classifier (described below) takes the independent components, each comprised of a spatial map and a time course, as input and labels each component as an EB, SA, CA artifact or a non-cardiac/non-blink/non-saccade independent component. From here onwards, the latter category will be referred to as the non-artifact (NA) label. Projection of only the NA components back onto sensor space reveals substantially cleaner sensor space signals (Fig. 2e, left) which, when used to reconstruct the brain source space activity, provides a more accurate estimate of actual brain activity (Fig. 2e, right).

2.3. Preprocessing

Data preprocessing steps included: down-sampling to 250 Hz, application of a notch filter to suppress line noise at 60 Hz and its first 2 harmonics, and band pass filtering to 1–100 Hz using the Brainstorm toolbox (Tadel et al., 2011).

2.4. Independent component analysis

The data was decomposed into 20 components via InfoMax ICA (Bell and Sejnowski, 1995). InfoMax is frequently used for MEG and is readily available in the Brainstorm toolbox. Each of these components consists of a pair of spatial maps and activation time courses, as shown in Fig. 3. The spatial map reveals the areas of magnetic influx (red) and outflux (blue) across the scalp while the time course indicates the temporal activation pattern of the spatial map during the MEG acquisition. Such pairs will be discussed in further detail in the subsequent sections. The number of components (20) is chosen for several reasons. First empirical analysis indicated that between 18 and 25 components yielded artifacts readily identifiable by our expert human readers, and there was unanimous consensus among the 4 expert raters that the artifacts were most identifiable using a 20 component decomposition. There is a tradeoff: at the low end, the raters noted that the artifacts are occasionally not well separated from non-artifact signal, while at the high end, artifacts more frequently split into multiple components. Multiple studies have identified between 8 and 14 canonical resting-state networks in resting-state functional MRI (rs-fMRI) (Beckmann et al., 2005; Giorgio et al., 2015; Heuvel and Hulshoff Pol, 2010; Smitha et al., 2009) and the standard Brain Nexus atlas includes 13 templates (Resting-State fMRI Templates – SCANlab, 2020). These networks have largely been shown to extend to electrophysiological neuroimaging, including MEG (Brookes et al., 2011; Coquelet et al., 2020; van Dyck et al., 2020). Thus, the choice of 20 ICA components presents a reasonable balance: it provides enough components to account for both the variability due to canonical biological networks and artifact sources while not dividing components into unrecognizable waveforms. In this study, the predictive models use both the spatial map and the time course components as inputs.

2.4.1. Ground truth labeling of ICA components

The component pairs from all subjects are independently classified as EB, SA, CA, or NA signal by 4 expert raters, with more than 23 years of experience in MEG data interpretation between them. If a component pair was not identically labeled by 3 or more raters, then it was flagged for discussion. A total of 52 of the 5880 component pairs were flagged. The raters then discussed and came to unanimous single label consensus for 39, while there remained a split decision for 10, with a split between artifacts and neuronal. These 10 were assigned the label NA, to encourage the subsequently trained predictive model to favor retaining signal

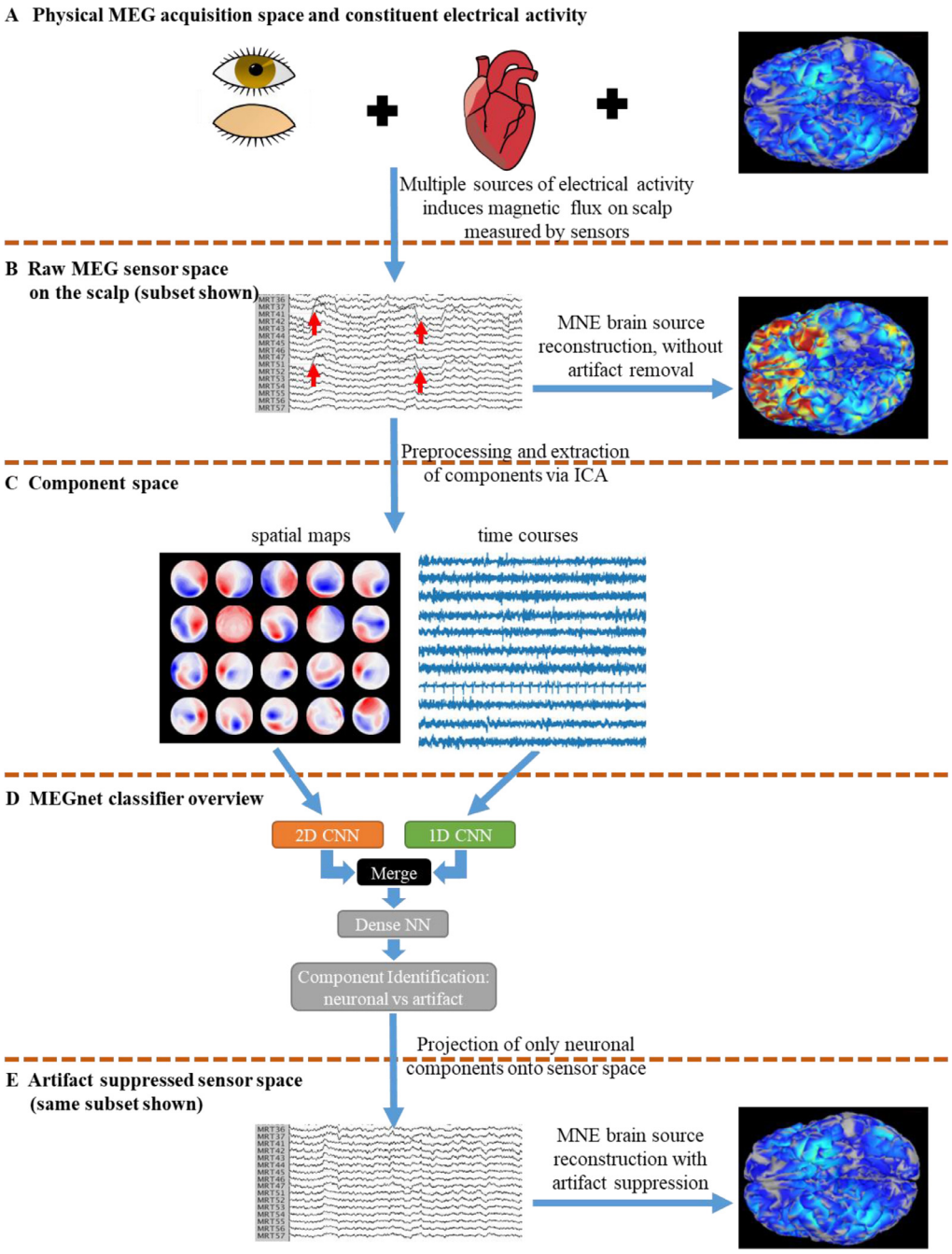

Fig. 2. The acquisition of MEG data and proposed post-processing pipeline with comparison to direct source reconstruction from sensor space time series. (A) Primary electrical activities in the physical acquisition space of the MEG scanner include: eye movements, particularly eye-blinks, cardiac activity, and electrical activity from neuronal firing (right). (B) Raw recorded sensor space of sensors near the scalp, with perturbation from blink artifacts indicated by red arrows (left) and a direct reconstruction in brain source space, without artifact removal (right). (C) ICA component extraction. (D) Overview of proposed MEGnet classifier to identify neuronal components. (E) Projection of components onto sensor space, with CA and EB artifacts removed, allows for a more faithful reconstruction of actual activity in brain source space (right).

in such cases. For the remaining 3 component pairs there was unanimous agreement that the component pairs contained both saccade and blink signal, these were subsequently labeled as saccade. Out of the total 5880 components, 4988 were labeled as NA, 285 EB, 183 SA, and 424 and CA. Representative examples of the component pairs from three subjects are shown in Fig. 4. The three components shown in the first row of Fig. 4b are illustrative of the inter-subject variation observed in the spatial maps of the cardiac artifact. The corresponding inter-subject variability in the time courses is shown in the top most panel of Fig. 4a. Subsequent rows in Fig. 4 show the variation across subjects in the spatial maps and time courses of the EB, SA, and NA signals.

2.4.2. Preparation of the 2D-Spatial maps

The preprocessing pipeline renders the spatial maps from ICA as topographic maps in the form of colored RGB images for ease of human interpretation, examples illustrated in Fig. 4a. The spatial maps are generated using the "2D disk" display of Brainstorm, which projects the flux information from the 3D arrangement of sensors near the scalp surface onto a standardized 2D circular space while minimizing distortion (Tadel et al., 2020.000Z). To reduce the input size, the 2D images are cropped to the bounding box containing just the disk, and has a final dimension of 120 pixels x 120 pixels x 3 color channels.

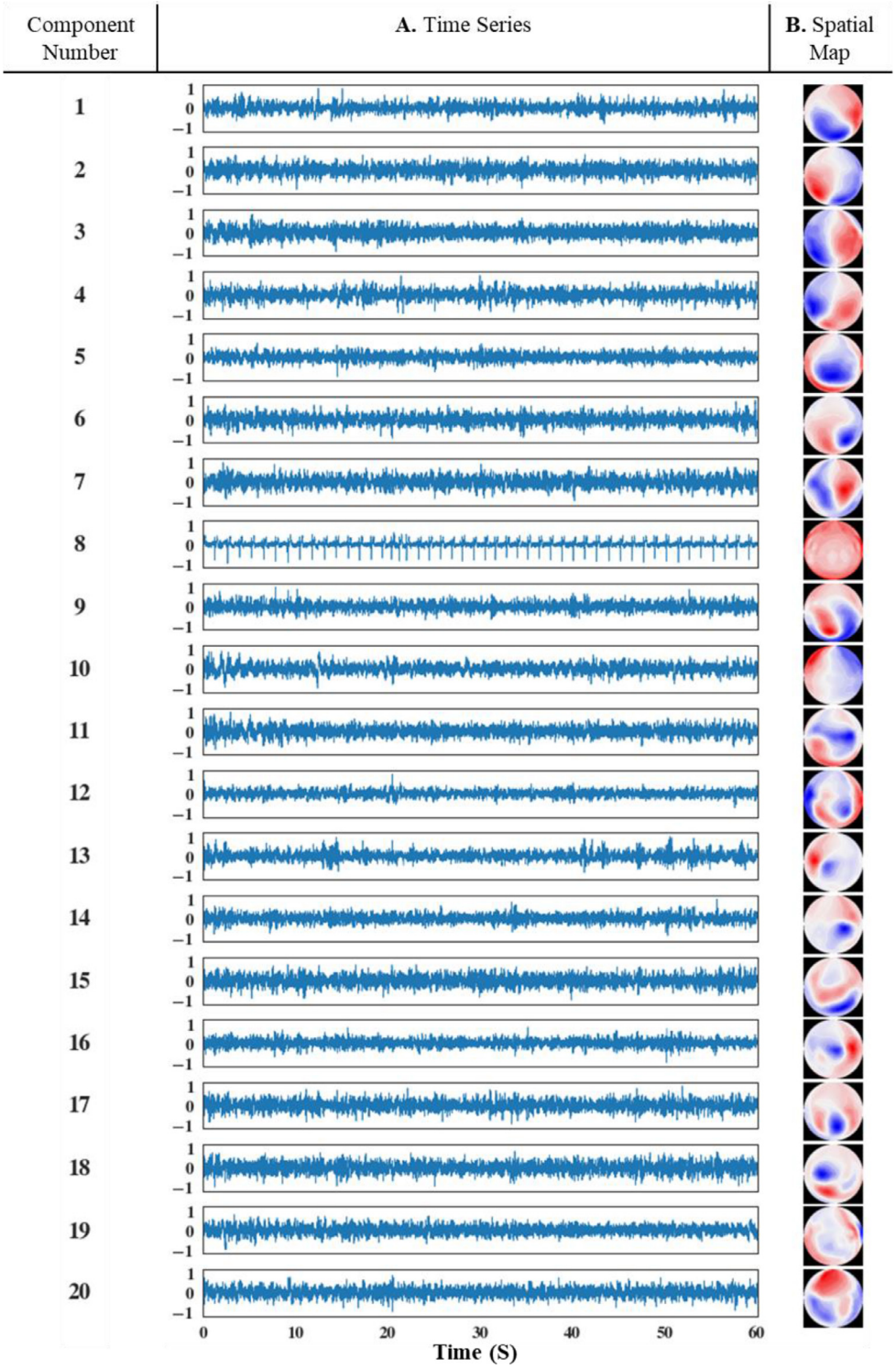

Fig. 3. Representative example of the ICA components extracted from a single subject. Each of the 20 components consists of a spatial map and a time course of map activation. (A) The 20 activation time courses of each spatial (B). The corresponding spatial maps capturing magnetic influx (red) and outflux (blue). The spatial map is projected into a 2D disk, and orientated as a top down view, with the subject's nose at the top.

2.4.3. Preparation of the 1D-time courses

In order to make the trained classifier capable of handling recordings of varying lengths, the time components are split into 60 second epochs (15,000 time-steps when sampled at 250 Hz) with a 15 second overlap, show in Fig. 5. In order to use all the data when the time series cannot be evenly split, the final 60 second epoch is taken as the last 60 s of the scan, and has a larger than 15 second overlap with the prior epoch. This approach, allows for *all acquired data* to be used both for training and testing the model. The 15 second overlap ensures that any predictive signal will be completely captured in at least one epoch without any edge effects. Additional information on the implementation for both training and testing described in Section 2.8. Sixty second epochs are used as the blink artifact has the largest period of the classified artifacts. With an average blink interval of about 20 s, a 60 second epoch will typically contain signal from at least 2 blinks.

2.5. Convolutional neural networks

Convolutional neural networks (CNN) have demonstrated remarkable success identifying real world objects in images in the Image-

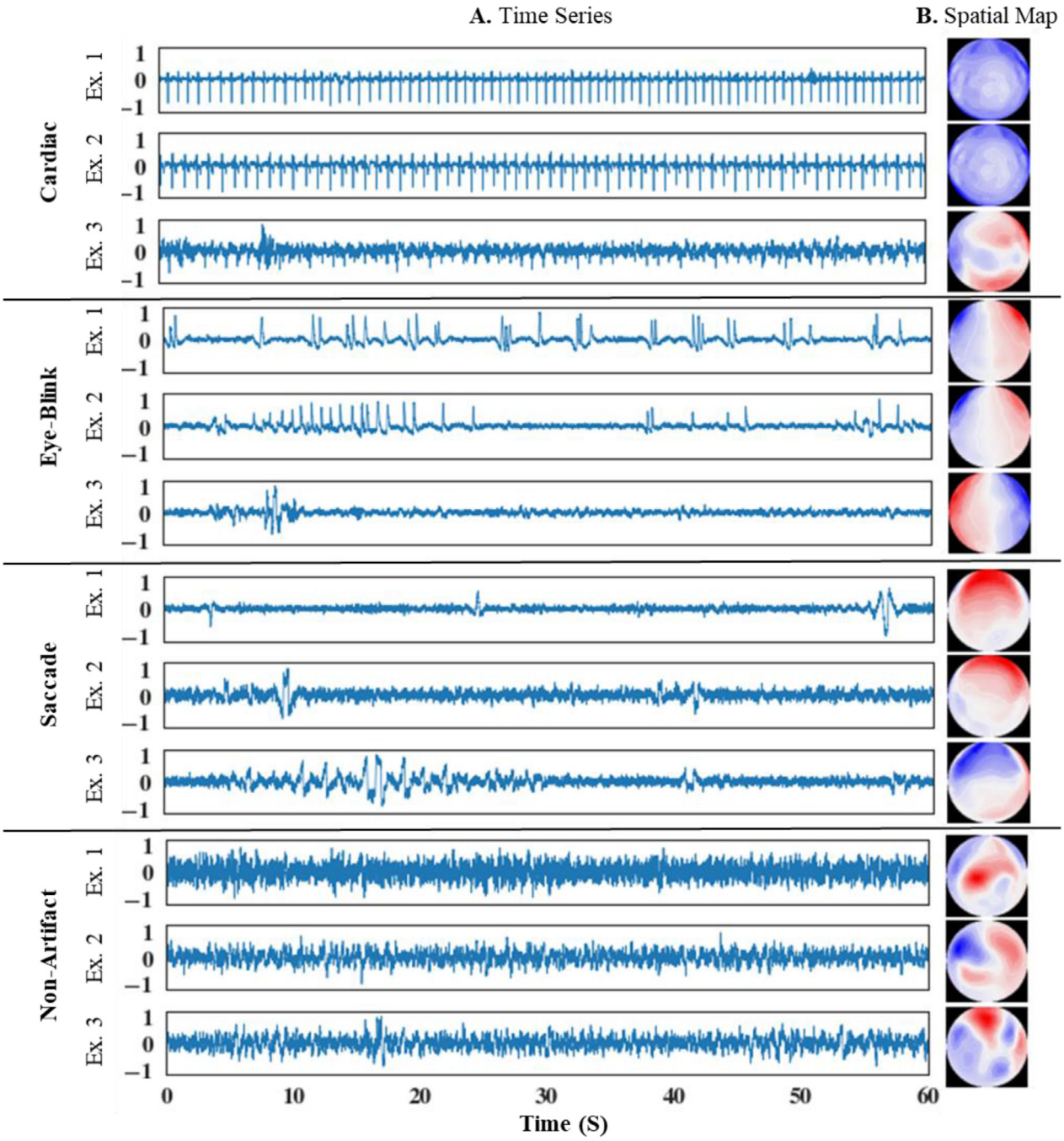

Fig. 4. Inter-subject variability in spatial maps and time courses is apparent in these four representative subjects. (A) Time courses for each signal category (B). Spatial maps from 4 subjects for each signal category.

net Large Scale Visual Recognition Challenge (Krizhevsky et al., 2017; Russakovsky et al., 2015; Simonyan and Zisserman, 2015). Such 2D CNNs (2D images with additional color channel) automatically learn the appropriate filters needed to accurately categorize the contents of a color image. Prior to the use of CNNs, the best algorithms used filters with manually crafted coefficients, which were applied to the images to extract local features. By 2016, through refinement of the CNN approach, the error rate surpassed human object recognition performance achieving an error rate of less than 3%. Inspired by these successes, the classifiers evaluated in this study employ combinations of CNNs.

2.6. Model construction

The overall structure of the proposed model's architecture consists of three subnetworks as illustrated in Fig. 2d. The general structure of the model entails a two-dimensional CNN to process the spatial maps, a one-dimensional CNN to process the time courses, and a dense feedforward network that merges via concatenation the latent representations learned by the two CNNs and outputs the predicted component class: eye-blink artifact, saccade artifact, cardiac artifact, or NA. Importantly the specific architectural design has been optimized through an extensive random search, which is described in the following section.

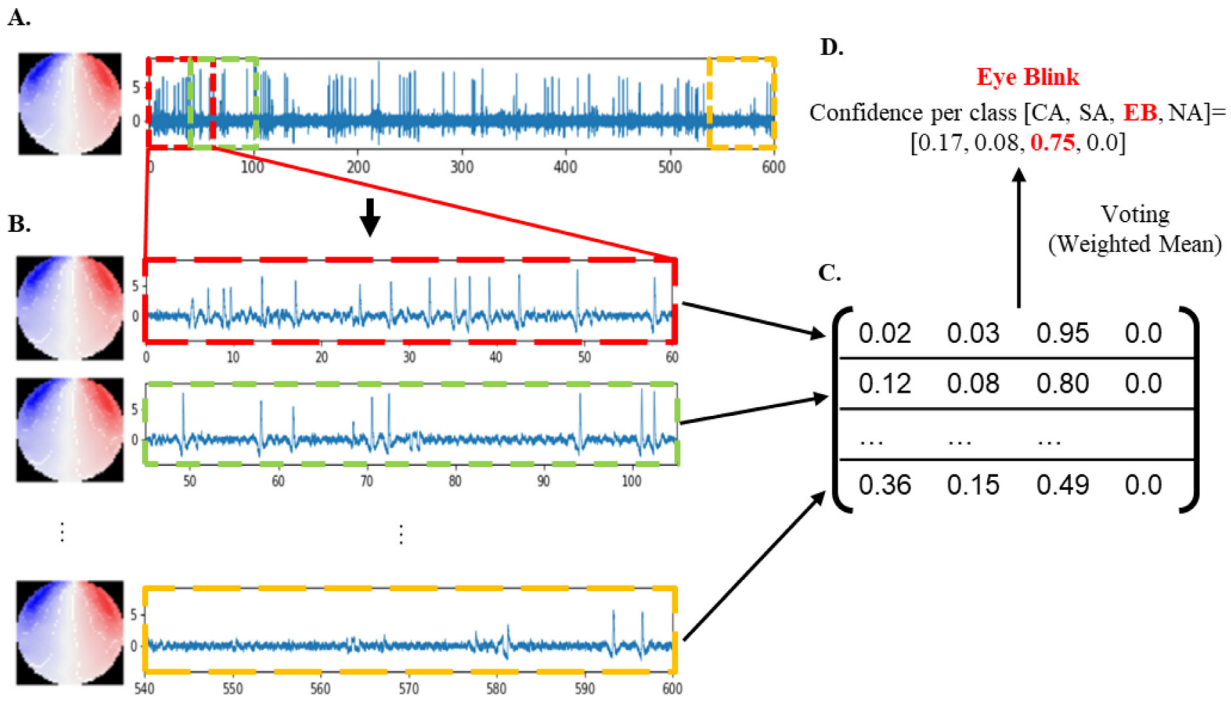

Fig. 5. Splitting time series into epochs and weighted voting. To provide a prediction on scans of varying length, the data is split into evenly sized 60 second epochs with a 15 second overlap, and voting is used for the final classification. A. The complete scan with a 10 min length. B. The complete scan is split into 60 second epochs, with corresponding spatial map. C. MEGnet is used to make a classification on each of the 60 second epochs. D. A weighted mean is used as a voting system to produce a final prediction for all of the data epochs.

2.7. Random search model optimization

There are many hyperparameters that must be chosen when constructing neural networks. In order to select optimal hyperparameters for the model's architecture, an extensive random search was conducted in which a total of 150 unique networks were constructed, trained, and evaluated using the random search (James Bergstra, 2012). To run the random search, a finite hyperparameter configuration space is defined and randomly sampled to define architectural configurations to evaluate. The hyperparameters of the configuration space and the ranges searched for each dimension are summarized in Table 2. The specific hyperparameters searched for each of our 3 subnetworks are described in the following 3 sections. In addition to these hyperparameters, which impact individual parts of the model architecture, there are 2 hyperparameters that are optimized that impact the entirety of each generated model. These include: (1) the *kernel weight initializer* that is selected from He normal, He uniform, Glorot normal or Glorot uniform, and, (2) the *activation function* that is selected between Parametric Rectified Linear Unit (PReLU) or Rectified Linear Units (ReLU). The order of batch normalization is subject to current debate, and therefore is also included in our hyperparameter search and the results are described in Section 4.2.

2.7.1. Search space of the spatial subnetwork

Specifically, for the spatial map 2D CNN subnetwork, the hyperparameters optimized included: the number of convolutional layers, the number of 2D convolutional filters for each layer, the kernel dimensions for each layer, whether to insert a maxpooling layer after each convolutional layer and whether batch normalization should be included after all of the convolutional layers in the spatial network. The *number of convolutional layers* was randomly drawn from a uniform distribution with a range of 1 to 10. The *number of filters* per convolutional layer was randomly drawn from a uniform distribution with a range of 1 to 64. The filter *kernel dimensions* were randomly selected from a uniform distribution with a range of 2 to 12, and was always square (height=width). The filters for the first layer span two dimensions in space and an additional dimension across color channel. The search space for the spatial network takes inspiration from both AlexNet and VGG-net. Similar to these networks the search space includes the options of max pooling layers and square kernel size spanning from 2×2 to 12×12. Batch normalization was also randomly chosen to be applied after each convolutional layer or not applied at all. After each individual convolutional network there was a 50% chance of adding a maxpooling layer with a window size of 2×2.

2.7.2. Search space of the temporal subnetwork

The configuration space searched for the time course 1D-CNN subnetwork included: the number of convolutional layers, the number of filters for each layer, the kernel size for each layer, whether to insert a maxpooling layer after each convolutional layer, and whether batch normalization should be included after all of the convolutional layers in the spatial network. The *number of convolutional* layers was randomly selected from a uniform distribution with a range of 1 to 10. The *number of filters* was drawn from a uniform distribution with a range of 1 to 64. The *kernel size* was drawn from a uniform distribution with a range of 2 to 16. The search space for the temporal network was similar to that of the spatial network, however the maximum kernel size was increased to 16 to ensure that for deeper networks, the last convolutional layer's receptive field size could cover two consecutive QRS complexes, even for subjects with very slow heartbeats of ~37.5 beats/min.

2.7.3. Search space of the dense feed forward neural network

To combine the outputs of both models, the latent representations learned by the 2D and 1D CNNs are combined via concatenation and are input into a dense feed forward network for classification. The search space for the dense network includes: (1) the number of layers ranging between 1 and 4, (2) the number of neurons ranging from 3 to 258, both of these hyperparameters are drawn from a uniform distribution, and (3) whether the input to the dense network should have batch normalization

Table 2
Defined hyperparameter search space and winning model's hyperparameters.

Spatial 2D Convolutional

Parameter	Range	Final Model Values
Number of 2D convolutional layers	1–10	8
Number of filters per layer	1–64	25,47,11,42,24,26,21,28
Kernel size (square) per layer	2–12	11,2,9,6,10,8,10,9
Max Pooling after each convolutional layer	T/F	F,T,T,F,F,T,F,T
Batch normalization after all convolution layers	T/F	T
Batch normalization before activation	T/F	F

Temporal 1D Convolutional

Parameter	Range	Final Model Values
Number of 1D convolutional layers	1–10	5
Number of filters per layer	1–64,	4,23,27,19,47
Kernel size per layer	2–16	5,4,12,9,8
Max Pooling after each convolutional layer	T/F	T,T,T,F,T
Batch normalization after all convolution layers	T/F	T
Batch normalization before activation	T/F	T

Dense Fully Connected Classifier

Parameter	Range	Final Model Values
Number of fully connected layers	1–4	3
Number of filters per layer	1–258	117,203,31
Dropout with rate of 0.5 after each layer	T/F	F,F,F,T
Batch normalize	T/F	T

Entire Model

Parameter	Options	Final Model Values
Activation function	ReLU, PReLU	PreLU
Kernel weight initializer	He normal, He uniform, Glorot normal, Glorot uniform	He uniform

applied. A dropout layer has the possibility of being added after the dense layers with a dropout rate of 0.5, similar to AlexNet and VGG-net.

2.8. Model training, selection, and final evaluation on held out test data

2.8.1. Model training through optimization of individual model weights

To optimize the weights of each individual model configuration the Adam optimizer (Kingma and Ba, 2014) was used. Adam was selected as it combines the desirable properties from 2 commonly used optimizers: RMSprop and AdaGrad by including both the first and second moments of the gradient. The categorical cross-entropy loss was selected as it outputs a probability over the set of classes for each component and this has been shown in the literature to produce high performance for multi-class classifiers (e.g. AlexNet and GoogleNet). Balancing per class was achieved by weighting the loss function by the ratio of each class. Each fold was trained for a maximum of 500 epochs (i.e. the number of iterations that the training data is used to update the models weights via back propagation). While training models for hyperparameter optimization, early stopping is employed to monitor the validation F_1 score and halt training when the F_1 score stops increasing. This helps ensure that models do not overfit to the training data.

2.8.2. Cross-validation based model selection

To compare performance across models, stratified group k-fold cross-validation is employed in which the components from a subject are grouped together such that they are in either the training or validation set, but not both. Using this approach, the 217 training scans are split into 10 folds with roughly 23 scans per fold. None of the 62 scans from the test set subjects are used for model selection. The splits were stratified for age, sex, site (source database), and scan type (resting-state, motor task, memory task, language task). Subsequent statistical testing confirmed that there were no statistically significant differences between the complete dataset and each split, Supplemental Table S1. As the temporal data is of varying length, the temporal data is split into 60-second epochs. Exact numbers of epochs for each train, validation, and test split is provided in Supplemental Table S2.

To select the winning model, the models are ranked according to the lower bound of the 95% confidence interval of the F_1 macro score across the 10 folds. The score for each of the four classes in our model is then calculated, where class $F_1 = (\frac{2*precision*recall}{precision + recall})$. Then, the mean of the four scores irrespective of the number of samples in each class is taken, also known as the F_1 macro score. The F_1 macro score allows each class to be weighted equally giving a fair account of the model performance, even in the presence of class imbalance. Additional performance metrics include the F_1 *micro* score and a confusion matrix. The confusion matrix includes performance for each class individually including sensitivity (true positive rate), false negative rate, positive predictive value, and false discovery rate.

2.8.3. Performance estimation on the test set

After selecting the winning model configuration, the model was trained on all training data, similarly split into 60 second epochs, and then evaluated on the held out test set of 1240 ICA component pairs from 62 test set scans. These 62 scans are not used for model training, hyperparameter optimization, or model selection. For final performance on classification of the 1240 ICA components (20 from each of the 62 scans) test data, the model is evaluated based on its performance labeling components (Fig. 5) as well as its performance labeling individual epochs within components. The final classification of each entire ICA component is formed from the weighted mean, over the posterior distribution of the models prediction for each class (Fig. 5C and 5D). The weighting is calculated by balancing the amount each time point contributes to the overall prediction, to ensure time points that occur in multiple epochs don't outweigh those in only one epoch.

2.9. Revealing what the model has learned

An important approach to gain insight into the abstraction learned by the proposed model, is to examine the components it labels correctly, with both high and low confidence. For each ICA component pair, the model outputs a confidence for each class (Fig. 5D), the class with the highest confidence is the predicted class. Those examples labeled with high confidence, are the ones that the model has learned well are canonical representations of the class, and contain features that are central to the label in its learned abstraction. Conversely, those example components, which the model predicts correctly but with lower confidence, are in the periphery of its learned abstraction. Results from such an analysis can be compared to what human experts might consider canonical features of each artifact class and neuronal component class.

Another important approach to gain insight into the abstraction learned by the proposed model, is to examine the importance or weight the model assigns to specific inputs (features). An approach to achieve this is called gradient class activation mapping (grad-CAM). This approach reveals which elements of an input feature vector are important for a the model to make a specific class prediction (Selvaraju et al., 2017). Grad-CAM uses the class-specific gradient information flowing into the final convolutional layer of a CNN to compute a localization map of the regions in the image important for making a classification. This method is broadly applicable to CNNs, including both the 1D time series and 2D spatial map sub-networks used in this work.

2.10. Ablation study

An ablation study was conducted to determine if both the time course and spatial maps are required to achieve maximum performance of the winning model. The model was split into two single input network models, one containing the layers from the spatial map subnetwork (Fig. 7 left subnetwork) and the other containing the layers from the time course subnetwork (Fig. 7 right subnetwork). For each single input model, the model was trained using the 10 fold cross validation data and, like the complete model, early stopping was used to determine the optimal number of training epochs. To provide an equivalent comparison as well as statistical significance between the ablation tests, each of the models' performance is measured and compared on the validation data across the 10 folds.

2.11. Analysis of the architectural search

The architectural search generates information about which hyperparameter combinations are likely to produce high and low performing models. To learn new lessons from this information, the performance of the top and bottom performing models is visualized to reveal what parameters they tended to have using kernel density plots. In particular, kernel density estimates (KDEs) are produced for the top and bottom 25% of models. In addition to the KDE plots, a contour plot of maximum model performance per hyperparameter combination is also generated to reveal a terrain map of model performance over hyperparameter space. Visualizing the complete 18 dimensional space is not feasible, however pairs of hyperparameters can be displayed as terrain maps. For this work, hyperparameter pairs are chosen that apply to the entire model, including activation, and number of layers, rather than hyperparameters like the number of filters per layer, which pertain to only a small part of the model.

2.12. Implementation of MEGNet

MEGnet is written in Python 3.7.10 (van Rossum and Drake Jr, 1995) using Keras 2.4.0 (Chollet and et. al, 2015) with Tensorflow 2.4.1 (Martín Abadi et al., 2015) as the backend for the machine learning models. The implementation further makes of use these python modules: numpy v1.19.2 (Harris et al., 2020), pandas v1.2.3 (McKinney, 2010), sklearn v0.24.1 (Fabian Pedregosa et al., 2011). Ray tune v1.2.0 (Liaw et al., 2018) is used for the hyperparameter optimization. The Brainstorm toolbox v3.1 (Tadel et al., 2011) is used for MEG preprocessing and ICA extraction.

3. Results

3.1. Ground truth labeling of ICA components: inter-observer agreement

To measure the inter-observer agreement between the expert raters, Fleiss' kappa and overall agreement percent are reported. Fleiss' kappa measures the degree of agreement above what would be expected by chance and ranges from less than zero to one. A kappa greater than 0.81 is often considered almost perfect agreement, and a kappa above 0.61 is substantial agreement. Overall agreement is calculated as the mean agreement across all raters for each ICA component rating. Inter-observer agreement of the 4 expert raters was very high with a Fleiss' kappa of 0.938 and an overall agreement of 97.5%. High agreement was also found on a per-class bases: for the class NA the Fleiss' kappa and overall agreement is 0.971 and 98.9% respectively, 0.87 and 95.1% for EB, 0.737 and 90.2% for SA, and 0.824 and 93.4% for CA.

3.2. Random search model optimization

The performance of the 150 models from the unbiased architecture search is illustrated in Fig. 6. Detailed performance results for each of the 150 models tested, are provided in Supplemental Table S3. A wide range in performance is attained across model configurations. Some models performed very well (Fig. 6 green), while others showed moderate ability to classify artifacts (yellow), and others demonstrated suboptimal performance (red). The figure inset provides a detailed comparison of the models that performed best. The model with the highest lower bound of the estimated F_1 macro score range over the validation folds is chosen as the winning (selected) model. The change in maximum performing model was also monitored during the search. When the maximum performance reached an asymptote, the search was stopped as the convergence indicated there is little further performance attainable by training additional models.

The architecture of the top performing model is shown in Fig. 7 and described in Table 2 (right column). In this model, the CNN subnetwork that processes the 2D spatial maps employs 8 convolution layers, 4 followed by max pooling (Fig. 7, left subnetwork). The 1D-CNN subnetwork processing the time course information (Fig. 7, right subnetwork) contains 5 convolution layers each, 4 followed by max pooling. Latent representations of the spatial maps and time courses are flattened, concatenated and used as inputs to the dense feed forward subnetwork. This subnetwork (Fig. 7, bottom) has three fully connected layers, the last layer having drop out applied while training, and terminated by an additional softmax layer that outputs the probabilities of each of the component class: eye-blink, saccade, cardiac artifact, or NA.

3.3. Performance estimation on the test set

The top performing model demonstrates very high performance on the held-out test set. This model performance was quantified 3 different ways. (1) First, it was computed at the whole component level, which is the intended mode for use. Here the model attained an overall classification accuracy of 98.87% and an F_1 macro score of 96.60% and an F_1 micro of 98.87%. The overall artifact sensitivity and specificity are 96.73% and 99.34% respectively. Complete details of the model's performance on the held out test data (not used during training or model optimization) is shown in the confusion matrix in Fig. 8, and summarized in Supplemental Fig. S1A. In Fig. 8, the first four rows are the actual classes of the target components, while the first four columns correspond to their predicted classes. In each cell the number indicates

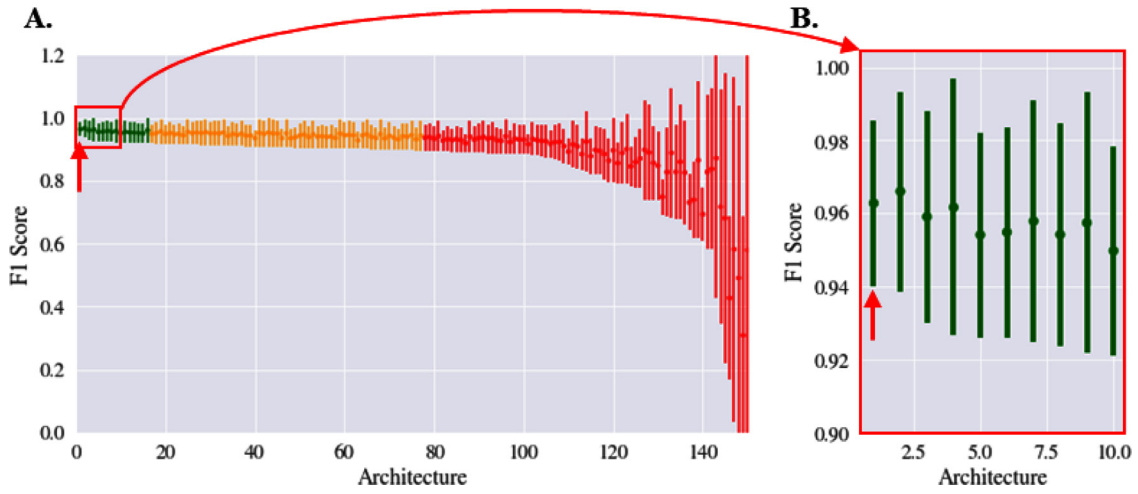

Fig.. 6. Random search over model architectures reveals wide performance variation. **A.** The performance for all 150 tested networks ordered by the lower bound of the 95% confidence interval (shown as a bar) of the F_1 macro score across the 10 fold cross validation. The architecture for the winning model is indicated with the red arrow. Models can be roughly categorized by performance. In red are models with low F_1 macro and/or high variance. Yellow highlighted models demonstrate suboptimal performance. **B.** Models highlighted in green demonstrate the best performing models.

the number of components. The right most column provides the sensitivity and false negative range (FNR) for each class, bottom row shows the positive predictive value (PPV) and false discovery rate for each class. The model achieves a sensitivity or true positive rate of 98.28%, 94.44%, 95.60%, and 99.34% for EB, SA, CA, and NA components, respectively. The model also obtains a PPV of 98.28%, 91.89%, 95.60%, and 99.43% for EB, SA, CA, and NA components, respectively. (2) Second, performance on individual epochs was computed and is detailed in Supplemental Fig. S1B where the model had an accuracy of 98.54%. (3) Third, model performance evaluated for each scan type, resting and task. This revealed that the model performs equivalently on both resting-state and task MEG. When the test data is split into resting-state and task the model achieves an accuracy, F_1 macro and F_1 micro of 98.57%, 96.04%, 98.57% on the resting-state data respectively, and 98.95%, 97.03%, 98.95% on the task-based MEG respectively. Confusion matrices for resting-state and task-based data are shown in Supplemental Fig. S1C and S1D respectively.

3.4. Revealing what the model has learned: What are the characteristics of components predicted correctly with high and low confidence?

In the proposed model, the softmax layer outputs the input component's class probabilities and the predicted class is the class with maximum probability. This probability can be considered the model's confidence in the predicted class, where a confident prediction has a probability of near 100%. The predictions for each component class are examined. First, for components classified correctly as *cardiac artifact* with high confidence, the spatial maps (Fig. 9) have small, gradual change in flux over the scalp and the temporal series contain a strong, regularly repeating signal in the frequency range of a human heart beat (~60bmp). Cardiac components predicted correctly with lower relative confidence contain more flux signal in the center of the spatial map and have a temporal signal with greater noise amplitude between the repeating signal peaks. This tends to agree with the human expert notion of a cardiac artifact, with predominantly signal near the center of the scalp spatial map and with a ~60 bpm frequency. Second, for components classified correctly as *eye-blink artifact*, the high confidence spatial maps contain strong influx and outflux signal bilaterally in the ocular regions and have time courses with characteristic trough waveforms (dips) that are indicative of an eye-blink, less regular than a heartbeat, and with a longer duration between the waveforms blinks than the heartbeat waveforms (Fig. 10). The lower confidence correct eye-blink components (e.g. the bottom row with 58.91% confidence) lack a smooth bilateral signal in the spatial maps. Third, the high confidence saccade artifacts (Fig. 11) show large signal in the time series with more irregular spacing indicative of saccades, and the spatial maps tend to have more symmetric flux with a large deviation by the ocular region. The lower confidence correctly predicted saccade components have less pronounced flux deviation across the ocular regions, or have more noise between saccade signal in the temporal component. Finally, for components classified correctly as NA with high confidence, the spatial maps do not have strong bilateral influx and outflux regions in the scalp periphery, but rather have strong signal fluctuation nearer to the center of the scalp. The NA time courses also show no regularly spaced peaks or troughs that could be indicative of heartbeats, or large isolated spikes in activity suggestive of eye-blinks (Fig. 12). Collectively, these results provide insight into the model's learned abstraction for each class. They show that the model reports a high confidence on inputs that clearly belong to a certain component classification and produce lower confidence on the harder inputs that would also be harder to classify by a human expert.

3.5. Revealing what the model has learned: What important features are learned from spatial maps and time courses?

The results of applying Grad-CAM to the temporal and spatial components are shown in Fig. 13. When the model is applied to a correctly predicted *cardiac component*, (Fig. 13, top row), the feature importance (red curve) peaks in unison with the heartbeats (spikes, blue curve). The feature importance in the spatial map shown by a black to green to yellow overlay, indicate that the model focused on the small, gradual change in flux across much of the scalp. When the model is applied to a correctly predicted *eye-blink* component, (Fig. 13, second row), the feature importance (red curve) peaks align perfectly with the signal troughs (blue curve) that are characteristic of eye-blinks in the time course. Meanwhile, the spatial map overlay shows a focus on both orbital lobes and the center of the scalp, likely identifying the characteristic high edge color contrast at the regions of the two orbits and relative flux between orbits and compared to the center of the scalp during eye-blinks. When applied to the correctly predicted *saccade* components, (Fig. 13, third row), the feature importance curve is elevated during the periods of high signal fluctuations associated with ocular movement. Feature importance of the spatial map, indicates that the model is focused on the consistent signal in the center of the map, as well as the gradient in the front ocular region. When the model is applied to a correctly predicted

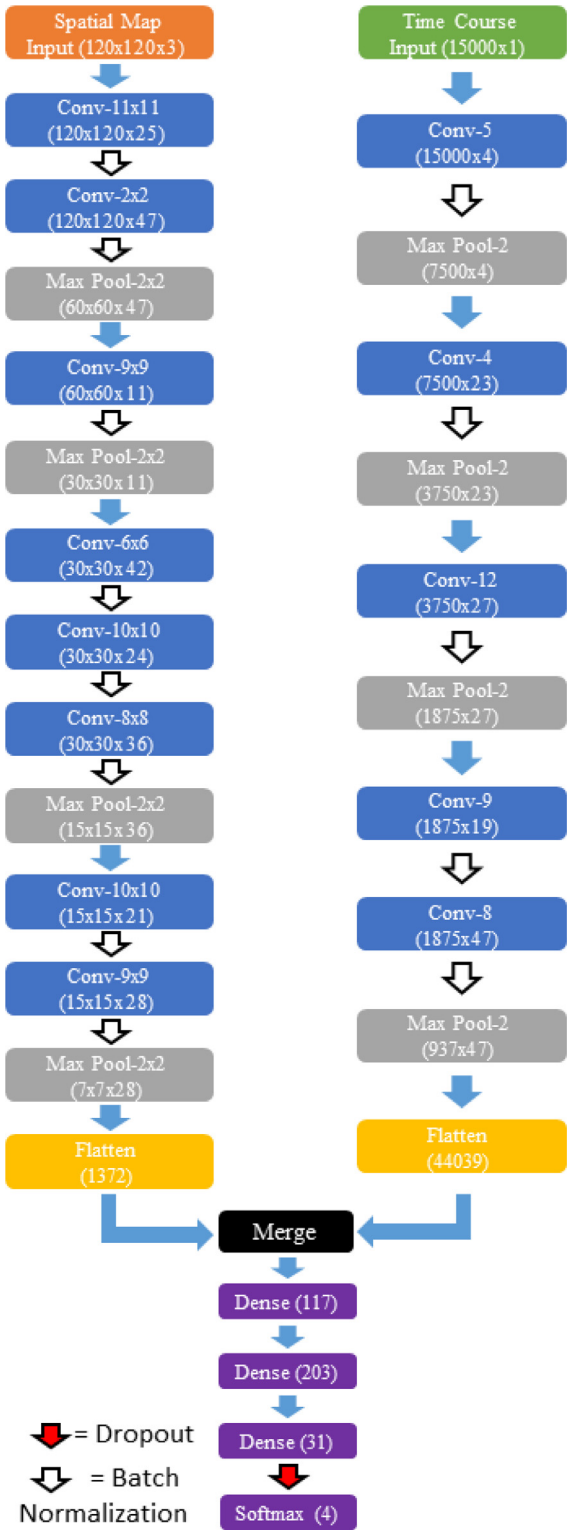

Fig. 7. Architecture of the best performing model. Overall, the spatial map and time course networks uses 8 and 5 convolutional layers respectively. 3 hidden layers are used in the dense merge network and drop out was used prior to the final layer. PReLU was used for the activation function along with He uniform initialization.

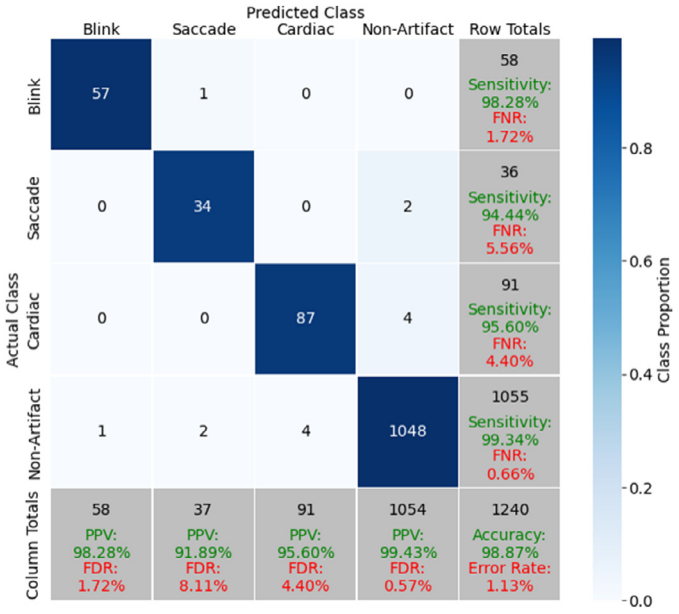

Fig. 8. Confusion matrix showing the winning model's performance on the held out test data. First four rows correspond to the actual (true) target classes, while the first four columns correspond to the model predicted classes. The bottom row shows the raw number and percentages of components predicted to belong to each class that are correctly (green) and incorrectly (red) classified called the precision (or positive predictive value) and false discovery rate, respectively. The right most column shows the percentages of all components belonging to each class that are correctly and incorrectly classified, called the sensitivity (or true positive rate) and false negative rate, respectively. The cell in the bottom right shows the raw component count and overall accuracy (green).

NA, (Fig. 13, bottom row) the feature importance (red curve) remains relatively high throughout the NA signal, which has characteristic high frequency oscillations throughout the time course. The spatial map overlay shows how the network correctly focuses on the center of the scalp in an area of high flux with a unique shape, which is typical for NA components. Taken together, these results suggest that the model has learned meaningful representations of the inputs and helps establish trust in the predictions made and abstractions learned by the model.

3.6. Ablation study

To determine if both the time course and spatial maps are required to achieve maximum performance of the winning model an ablation study was conducted. As shown in Fig. 14, the model using only time course information achieved a mean F_1 macro score of 80.1% with a standard deviation of 0.81%. Meanwhile, the model using only spatial map information worked statistically significantly better achieving a mean F_1 macro of 89.05% with a standard deviation of 1.62%. However, the final model that uses both the spatial map and time course inputs outperforms both single input models, achieving a mean F_1 macro of 96.3% with a standard deviation of 1.16%. These differences are significant with a p value < 0.00001. These results indicate that the spatial map and time course inputs contain complementary information and both contribute to the overall performance of the proposed model.

3.7. Analysis of the architectural search

The last 3 sections revealed insights into what the highest-performing model has learned and that there is a need to combine spatial and temporal information to obtain maximum performance. The extensive unbiased architecture search of 150 models also reveals insight into the hyperparameter configurations that tend to achieve high artifact classification performance.

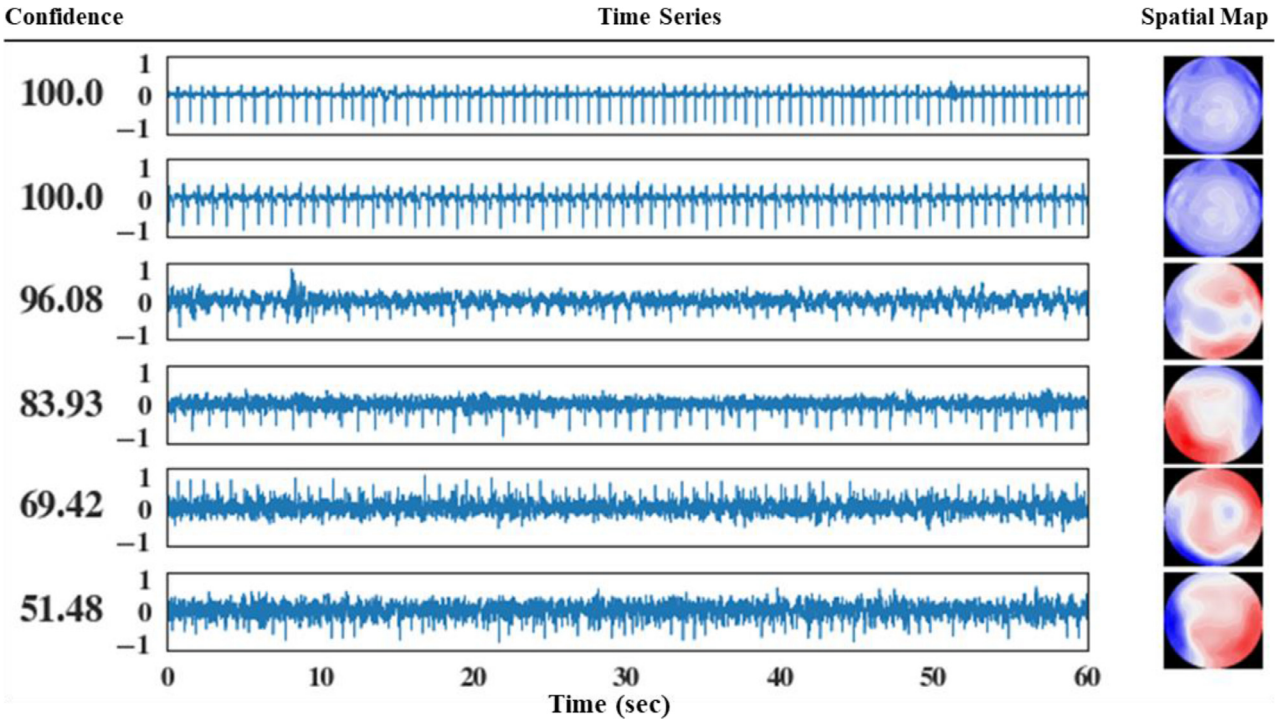

Fig. 9. Examples of correctly predicted cardiac artifact components, ordered from high confidence (top) to lower confidence (bottom).

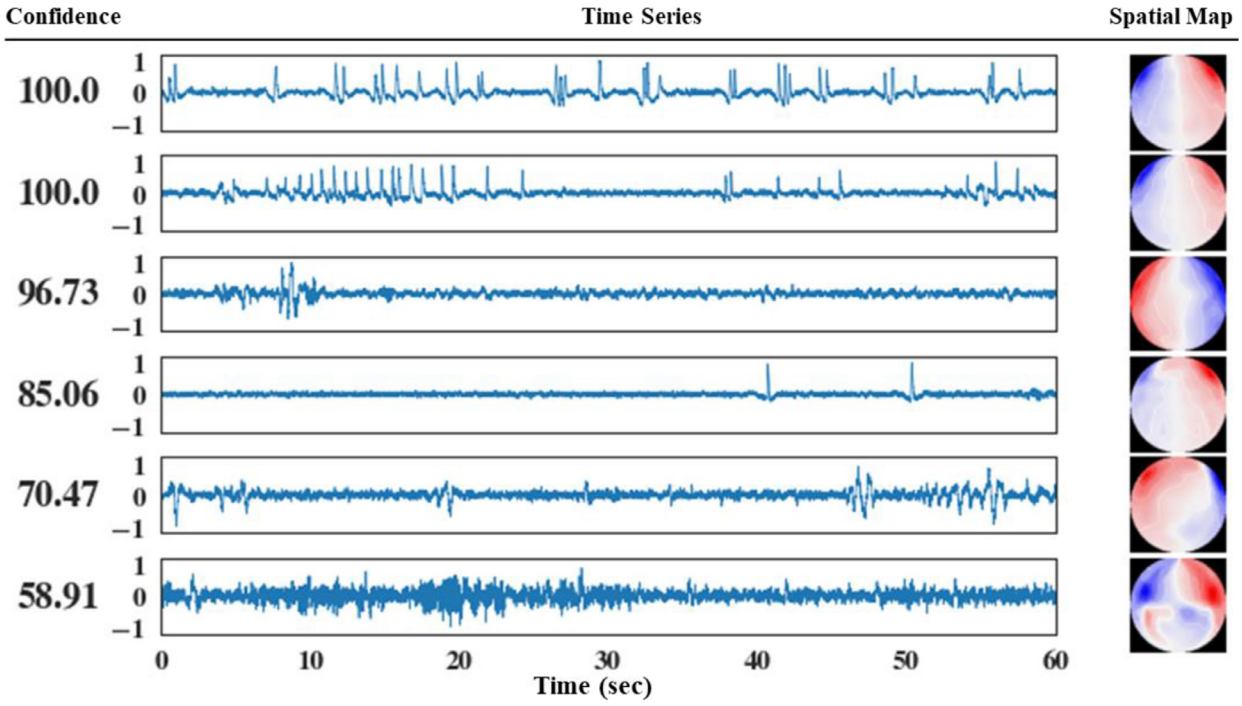

Fig. 10. Examples of correctly predicted eye-blink artifact components, ordered from high confidence (top) to lower confidence (bottom).

One way that this can be achieved is by visualizing the performance of the top and bottom performing models and examining what parameters they tended to have using kernel density plots. This reveals that there is no one ideal configuration, but rather a set of configurations that perform very well and it reveals which hyperparameter combinations tend to comprise good performing models. In particular, the highest and lowest performing 25% of the total 150 models, have been selected and the density plots for these two groups of models is shown in Fig. 15 as the high (green) and low (red) performing surfaces. Since the full hyperparameter space has 18 dimensions, for visualization, each plot shows two hyperparameters. In the hyperparameter subspace spanned by the number of layers in the spatial and temporal subnetworks (Fig. 15a), a few high performing (green) peaks are evident. *High performing networks typically have 2–4 or 6–8 convolutional layers in the spatial subnetwork and 5, 6, or 8 convolutional layers in temporal subnetwork*. For the hyperparameter subspace spanned by batch normalization in the temporal and

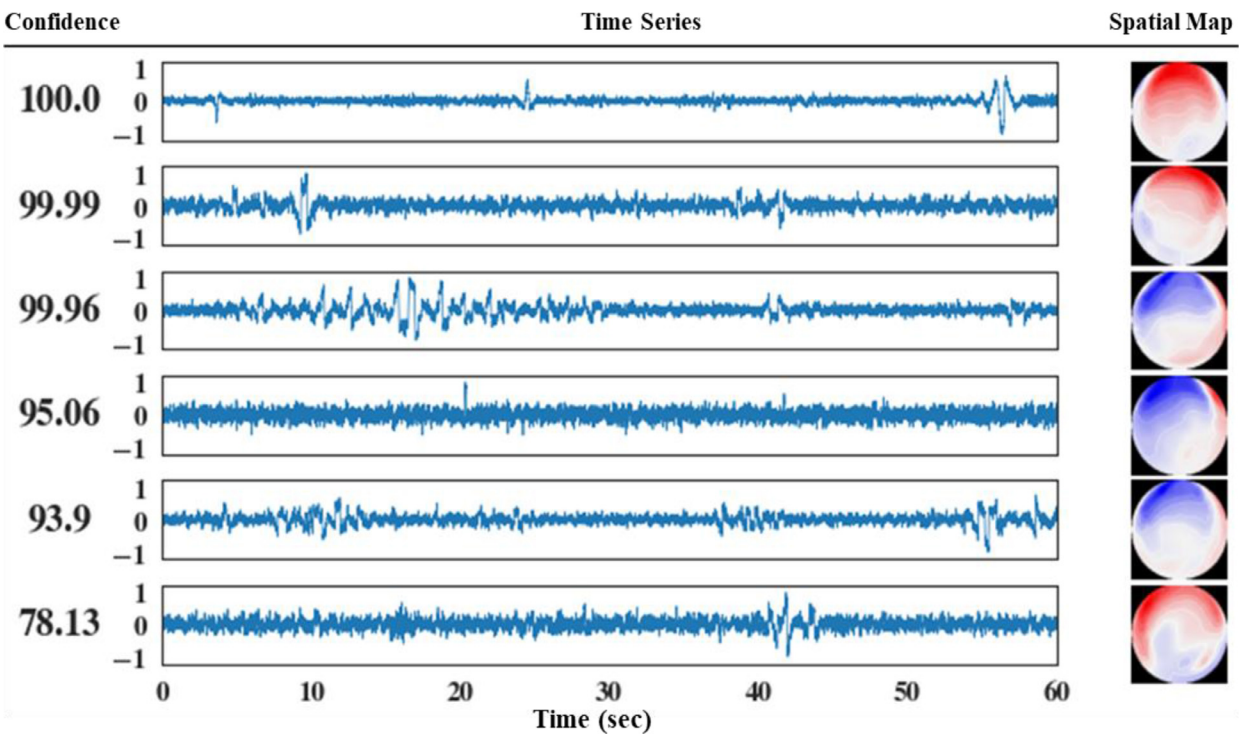

Fig. 11. Examples of correctly predicted saccade artifact components, ordered from high confidence (top) to lower confidence (bottom).

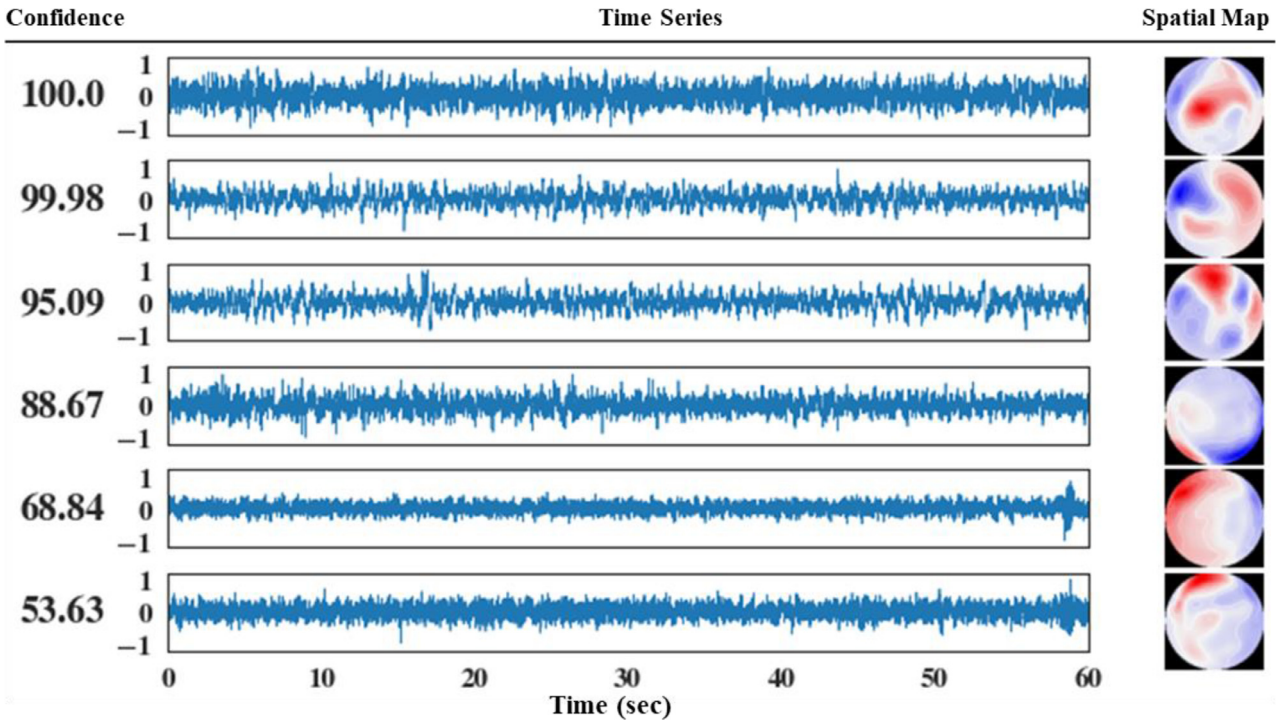

Fig. 12. Correctly predicted NA components are ordered from high confidence (top) to lower confidence (bottom).

spatial subnetworks, *there is a preference (green peak) to use normalization in the spatial subnetwork*, but not in the temporal subnetwork (Fig. 15b). Finally, for the hyperparameter subspace spanned by the number of layers in the dense neural network and the activation used for the network *there is a preference for 2 dense layers* (Fig. 15c).

KDEs indicate a preference of specific hyperparameters rather than the absolute performance attained for each hyperparameter combination. Another way to uncover insight into preferred hyperparameter configurations, is to visualize the highest performance attained at each point in hyperparameter space (Fig. 16). For this visualization, the 2D subspace (pair of hyperparameters) with highest performance variance over its axes was chosen. This more nuanced, perhaps more noisy view, confirms that there are several high performing regions in hyperparameter space (green regions) as well as several suboptimal configuration regions (red regions), where even the best performing model's with the hyperparameter combination did not perform well.

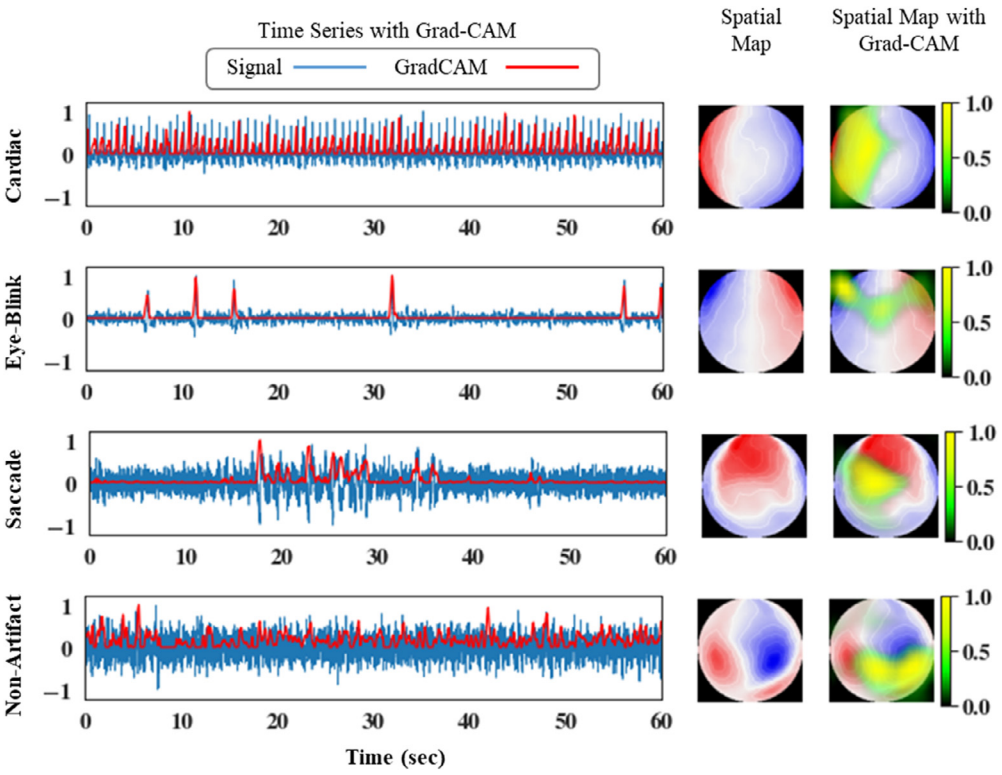

Fig. 13. Grad-CAM on an input from each class that the model predicts with high confidence.

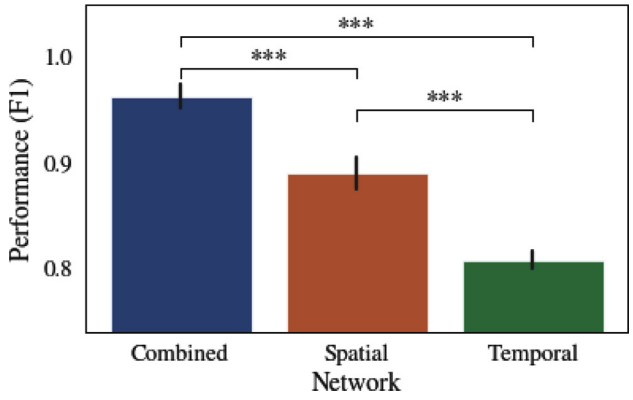

Fig. 14. Performance of the combined, spatial only and temporal only network across the k fold validation data. Early stopping was used to ensure each model did not overfit. Error bars show 1 standard deviation. All p values are highly significant at 3.99e-8, 2.32e-6, 3.71e-11 for spatial vs temporal, spatial vs both and temporal vs both respectively. P-values are calculated using a paired *t*-test. The highest performing model used both the spatial and temporal inputs.

4. Discussion

Based on our 4 expert raters, human expert inter-observer agreement is estimated to be 97.5% percent with a Fleiss' kappa of 0.938, thus the winning model achieves performance at or very nearly the same level as the human experts, and it does so regardless of whether the MEG data is resting-state or task-based. Of all the detected artifacts, saccades appear to be the hardest artifact to identify. This is reflected in slightly lower inter-human expert agreement and model performance, though the model does achieve good performance on saccade artifacts (sensitivity of 94.4% and PPV of 91.89%). Further, the analysis shows that the artifact detection by integrating voting across from the multiple epochs per component (Fig. 8) performs better in almost every single metric than single epoch prediction (Fig. S1B), suggesting that using the complete component time series is preferable for prediction than a single epoch.

Achieving such a high performance level, suggests that the method may be used to obviate the requirement for human experts to identify component artifacts. As the model performs well on the on the held out test set, it can be expected to work well on MEG scans of people belonging to either sex, and a broad range of ages (9 through 73 years) and on task and resting-state. The proposed model facilitates the use of MEG for large research studies and clinical applications, where a human expert may not be available or when the numbers of subjects is large, making human labeling problematic. The performance of the selected model compares favorably to the most closely related published works, which are further described in Section 4.1. Section 4.2 discusses the optimal ordering of batch normalization and activation, while Section 4.3 discusses limitations of this study.

4.1. Comparison to related work

To date there has been limited research into the automation of artifact removal in MEG without the use of supplementary electrodes such as EOG and ECG. To provide a most commensurable comparison, this section focuses on studies that do not use EOG and ECG electrodes. Three other papers, (Croce et al., 2019; Duan et al., 2013; Hasasneh et al., 2018) also aimed to remove artifacts from the ICA components. An overview comparing our work to those discussed here is presented in Table 3.

Duan et al. employed a support vector machine (SVM) that was trained with five manually selected features (probability density, kurtosis, spectral entropy, fractal dimension, and central moment of frequency) extracted from the time courses from ICA. Hasasneh et al. and Croce et al. applied multi-input deep learning networks that are similar to the models in this work. Duan et al., Hasasneh et al. and Croce et al. all report a cross-validation (*without* held-out test set).

Table 3
A comparison of related work in automated MEG artifact detection. MEGnet generally outperforms the other models. Performance of MEGnet is conservatively measured on unbiased, held out test data, whereas other models' typically report performance on validation data. Garg et al., 2017b is omitted as it classifies only eye blink artifacts, whereas all other methods classify both eye blink and cardiac artifacts. NR = Not Reported, Acc. = Accuracy, Sens. = Sensitivity, Spec. = Specificity.

	Type of Predictive Model	Data				Performance		
		Subjects (N)	ICA components (N)	Handled Scan Types	Ground truth method	Acc.	Sens.	Spec.
Duan et al.	SVM	10	956	Resting State	"manual inspection"	97.41%	92.01%	**99.65%**
Hasasneh et al.	Multi-input Deep Neural Network	48	1632	**Task-based and resting state**	Independent methods and "visual inspection"	94.4%	91.8%	97.4%
Garg et al., 2017a	Single Input CNN	49	980	Resting State	Single rater	95.86%	79.6%	98.2%
Croce et al.	Multi-input Deep Neural Network	NR	4749	**Task-based and resting state**	"trained experts"	95.5%	NR	NR
MEGnet	Multi-input Deep Neural Network	**217**	**49,100**	**Task-based and resting state**	**Independently by 4 experts**	**98.95%**	**96.74%**	**99.34%**

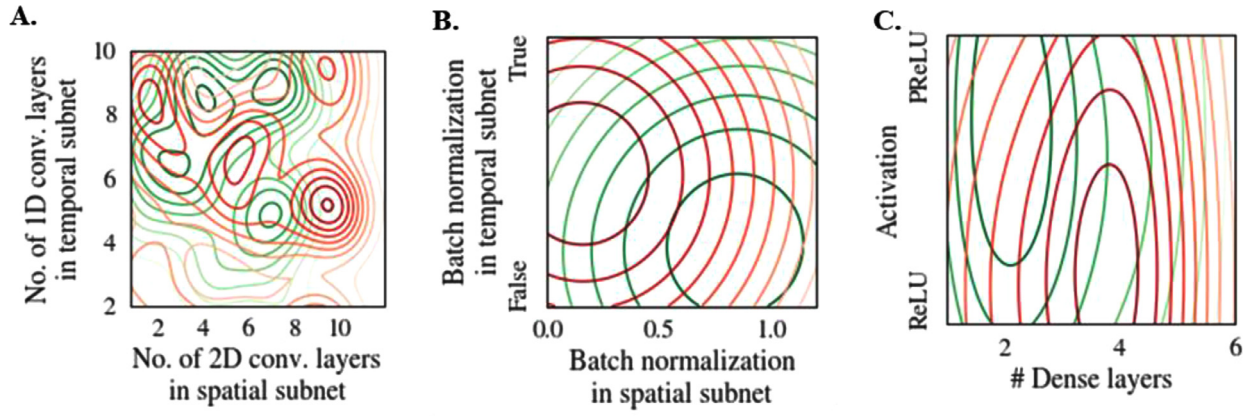

Fig. 15. Hyperparameters that tend to produce high and low performance. Kernel density plots of models with the top 25% and bottom 25% F_1 macro performance are shown in green and red respectively. A. For the number of layers in the temporal network and spatial network, four high performing (green) peaks are evident. B. For batch normalization in the temporal and spatial subnetworks, there is a preference (green peak) to use normalization in the spatial subnetwork. C. Among the number of layers in the dense neural network and the activation used for the network, there is a preference for 2 dense layers, and the opposite for 4 dense layers.

Notable in this comparison is that several other studies have used small sample sizes, which can be a significant impediment to reliable performance estimation. This study uses an extensive cohort with 217 subjects, with a broad age distribution of 9 years to 73 years, and a mixture of males and females.[1] Overall, the MEGnet model proposed in this work demonstrates higher performance including *98.95% accuracy, 96.74% sensitivity and 99.34%* specificity (Supplemental Fig. S1A). The proposed model outperforms previously proposed models in all metrics, except for specificity by Duan et al. which reported 99.65%, however this is within 0.31% of the proposed MEGnet model and Duan et al. report cross-validation performance which tends to be inflated compared to the more rigorous and conservative held out test performance, that is used in our investigation. The higher performance of the proposed model herein is likely due to several factors: (1) an extensive, unbiased hyperparameter optimization (Section 2.7), (2) the training upon data from multiple datasets, (3) and training upon a larger number of total subjects. We suggest that a model learned from multiple sites' data can help the model generalize well since the model has already learned information obtained from different technicians and acqusition protocols. In addition, the model proposed in this work is trained and tested on multiple types of scans including resting-state and 3 different tasks, this helps ensure that the model will generalize well reguardless of the type of scan. In comparison, works such as Duan et al. reports are based on a single scan type. Finally, in comparison to the work by Hasasneh et al. and Croce et al., this study provides insights to what the proposed MEGNet model has learned, giving further assurances that the proposed model will generalize well.

Our previous work (Garg et al., 2017a, 2017b), also removes artifacts. In those works, we separately used time course and spatial map components to identify artifacts. In comparison, this work expands on our previous work in numerous ways. The present work uses a multi-input deep neural network to extract features from both the spatial and the temporal components, learns to integrate that information optimally, and demonstrates increased predictive accuracy over a much broader test set.

Other, related work focuses on the identification of bad channels (sensors) in MEG acquisition. Notably, Autoreject (Jas et al., 2017) is one such bad channel rejection approach. We note that this work is distinct from and complementary to ours. Jas et al. identify and remove poorly performing *sensor information* from MEG data, while explicitly indicating that a complementary approach is needed to remove *non-neurological physiological artifacts.* Though not explored in this manuscript, the integration of Autoreject and MEGnet presents an interesting research direction to suppress both erroneous sensors and well as remove non-neurological signal sources. While successful at reducing sensor artifacts, the authors of Autoreject specifically note that their model does not completely remove biological artifacts, and that ICA methods "naturally supplement autoreject … [as they] extract and subsequently project out signal subspaces governed by physiological artifacts such as muscular, cardiac and ocular artifacts". This work comple-

[1] Duan et al. used 10 subjects roughly between 4-6 years old. Hasaneh and Croce et al. did not detail demographics.

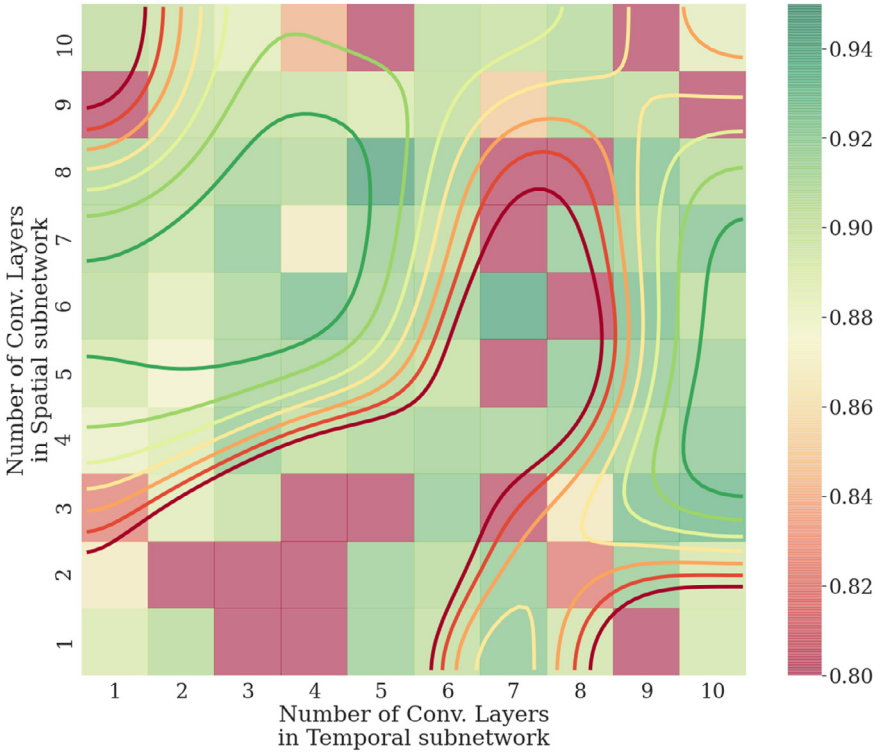

Fig. 16. Performance landscape within the hyperparameter subspace consisting of the number of layers in the spatial subnetwork and the number of layers in the temporal network. Maximum performance across the tested models is shown in each pixel entry. Several better performing configurations (greener regions) are evident. A contour plot that interpolates and smooths raw performance helps intuit these regions. The performance of each pixel is measured corresponding best performing model's lower 95% confidence interval of the F_1, measured on the validation data over the 10 fold cross validation. The final models parameters are indicated by the blue box.

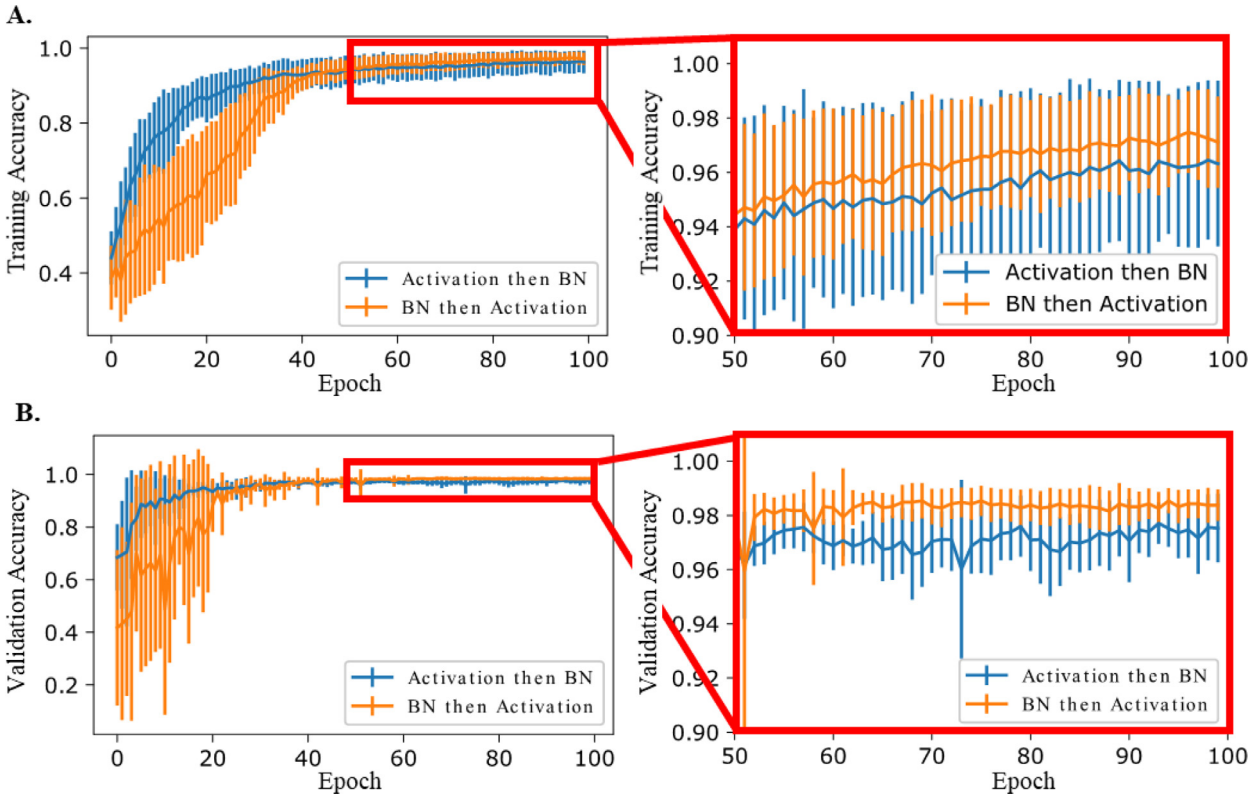

Fig. 17. The plots here show the importance of including the order of activation and batch normalization in the random search. The blue line represents the performance per epoch when the model has activation before batch normalization (BN) and the orange represents the model having batch normalization before activation. This was trained using 10 fold cross validation, and did not include the test data. Each point is the mean with error bars of one standard deviation across the 10 folds. A. The accuracy of the model per epoch on the training data. B. The accuracy of the model per epoch on the validation data.

ments Autoreject by automating the accurate identification and removal of these artifacts

4.2. What is the optimal ordering of batch normalization and activation?

Since there is open scientific debate regarding whether batch normalization should be performed before or after the application of the activation function in each convolutional layer, an experiment was run to determine the effect of the ordering. It consisted of creating two models each of the same dual input subnetwork form illustrated in Fig. 2d. In one, model batch normalization was performed before the activation layer while in the other, it was performed afterwards. Both models were then trained identically and the training and validation performance was computed using 10-fold cross validation (Fig. 17). These results suggest that there is no clearly superior ordering. When activation precedes batch normalization (blue curve), learning is faster, while when the opposite ordering is used (orange curve), the final performance is slightly better. Since these results may be architecture dependent, the ordering was included in the random search of Section 2.7. For the final model, shown in Fig. 7, batch normalization preceded activation in the temporal network, and followed activation in the spatial network, and is not used in the merge network.

4.3. Limitations

There are several limitations in this study. First, this study does not include patients with abnormal P-QRS-T patterns such as patients with a heart arrhythmia. Second, this study does not include patients with neurological disease, such as myasthenia gravis or patients with eye pathologies and in these populations abnormal eye-blinks can occur. Third, the final model was trained on subjects aged 9–73 years, but not on very young children and infants using specialized infant-MEG scanners; therefore, it may not detect artifacts in subjects with ages substantially lower than 9 years with the same high performance. Additionally, while this model was trained on 3 different tasks including sensory motor task, a working memory task, and a language processing, and thus will likely perform will on other tasks, performance on other tasks has not been tested thus cannot be guaranteed. However, the proposed model training approach includes a comprehensive architectural search that is fully automated. Therefore, with additional data spanning such cases the approach could be readily adapted. Also, we target identification of the most prolific and problematic cardiac, eye blink and saccade artifacts. We note that there are other artifacts, which could also be targeted. With an appropriately labeled dataset, we expect the model could be trained to identify such artifacts. To support such extensions, full source code is being made publically available. Finally, our ICA components are extracted using Brainstorm. Preprocessing on other software packages or with different processing steps (Section 2.3 and 2.4) could affect the performance of the model, however, brainstorm was selected for this work as it is widely used, reliable, and open source.

The proposed approach achieves human expert level performance, which suggests that it may be suitable for the analysis of other functional neuroimaging scenarios, particularly those in which ICA is already a commonly used preprocessing step, including: fMRI, EEG, and fNIRS. In these modalities, ICA can also yield time course and spatial map components much like those that the proposed model processes for MEG. Adaptation of the proposed model for fMRI would require extension to 3D, but the classification task would otherwise have multiple similarities.

5. Conclusion

MEG is a rapidly growing functional neuroimaging modality that has the potential to facilitate diagnoses and prognoses in a wide range of neurodegenerative diseases, psychological disorders and developmental disorders. It is already being used clinically for pre-surgical planning in epilepsy, brain tumors and other indications requiring brain resection. More recently MEG has shown promise to discriminate neurodegenerative disorders (Guillon et al., 2017; Nakamura et al., 2018; Olde Dubbelink et al., 2014), neurodevelopmental disorders (Kasturi Barik et al., 2020; Monge et al., 2015) and psychological disorders (Crunelli et al., 2020; Wang et al., 2019). To obtain the most useful signals from MEG, artifact identification and suppression is vital since these artifacts can corrupt large portions of the signal in brain space reconstructions. This work provides multiple contributions to the field of MEG neuroimaging data analysis. *First*, this paper proposes an artifact classification approach that combines multivariate decomposition with a deep learning multi-subnetwork model that fully automates artifact separation and detection directly in MEG data without the need for complicated patient setup procedures that use EOG (electrooculography) or ECG (electrocardiography). *Second*, compared to the published literature, the proposed model achieves new state of the art accuracy detecting MEG artifacts with 98.95% accuracy, 96.74% sensitivity and 99.34% specificity. The model achieves this expert human level artifact classification performance across a wide spectrum of subject ages: 9–73 years old. *Third*, this work utilizes a computational, unbiased model selection procedure ensuring that the proposed model is well suited to the artifact classification task. *Fourth,* this research also reveals insights about suitable candidate architectures from the unbiased model search, as well as the features and abstractions learned by the top performing model. *Fifth,* the study demonstrates that spatial maps and time courses contain complementary information and therefore need to be combined to achieve a top performing classifier. *Lastly*, the proposed method is fully automated, requiring no user input which facilitates automated MEG processing for clinical and research use and supports it adaptation for additional domains.

Data and code availability statement

To facilitate reuse and extension, we are pleased to provide full source code for the proposed approach at: https://github.com/DeepLearningForPrecisionHealthLab/MegNET_2020 The McGill OMEGA dataset is publically available at the following webpage: https://www.mcgill.ca/bic/resources/omega. The HCP dataset is publically available at the following webpage: https://www.humanconnectome.org/software/hcp-meg-pipelines.

Credit author statement

AM, AHT, PG, JAM and ED contributed to the conception of the experiments. AHT, AM, and PG contributed to the methodology, software implementation, execution of experiments, validation, and formal analysis, and interpretation of results. ED, RG, AP, LB contributed to data curation including expert rating. GM, BW contributed to data curation and computing resources. CTW, JDS, JAM, ED contributed to data acquisition and interpretation of the results. AM, AHT and PG wrote the manuscript draft with review and critique from all other authors.

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgments

Support for this research was provided by the Lyda Hill Foundation (AM), NIH NIA R01AG059288 (AM, AT), NIH R01NS082453 (JAM, JS), and R01NS091602 (JAM, CW, JS). This material is also based upon work supported by the NSF Graduate Research Fellowship under Grant #DGE-0907738. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the NSF.

We thank Julia Evans, Michael Rugg, and Satwik Rajaram and the reviewers from NeuroImage for their feedback during the writing of this manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2021.118402.

References

Beckmann, C.F., DeLuca, M., Devlin, J.T., Smith, S.M., 2005. Investigations into resting-state connectivity using independent component analysis. Philosophical transactions of the Royal Society of London. Series B, Biol. Sci. 360, 1001–1013. doi:10.1098/rstb.2005.1634.

Bell, A.J., Sejnowski, T.J., 1995. An information-maximization approach to blind separation and blind deconvolution. Neural Comput. 7, 1129–1159.

Bellec, P., Rosa-Neto, P., Lyttelton, O.C., Benali, H., Evans, A.C., 2010. Multi-level bootstrap analysis of stable clusters in resting-state fMRI. Neuroimage 51, 1126–1139. doi:10.1016/j.neuroimage.2010.02.082.

Breuer, L., Dammers, J., Roberts, T.P.L., Shah, N.J., 2014. Ocular and cardiac artifact rejection for real-time analysis in MEG. J. Neurosci. Methods 233, 105–114.

Brookes, M.J., Woolrich, M., Luckhoo, H., Price, D., Hale, J.R., Stephenson, M.C., Barnes, G.R., Smith, S.M., Morris, P.G., 2011. Investigating the electrophysiological basis of resting state networks using magnetoencephalography. Proc. Natl. Acad. Sci. U.S.A. 108, 16783–16788. doi:10.1073/pnas.1112685108.

Buzsáki, G., Anastassiou, C.A., Koch, C., 2012. The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes. Nat. Rev. Neurosci. 13, 407–420.

Chollet, F., et. al, 2015. Keras.

Coquelet, N., Tiège, X.de, Destoky, F., Roshchupkina, L., Bourguignon, M., Goldman, S., Peigneux, P., Wens, V., 2020. Comparing MEG and high-density EEG for intrinsic functional connectivity mapping. Neuroimage 210, 116556. doi:10.1016/j.neuroimage.2020.116556.

Criswell, E., Cram, J.R., 2011. Cram's Introduction to Surface Electromyography, 2nd ed. Jones and Bartlett, Sudbury, Mass. 1 online resource.

Croce, P., Zappasodi, F., Marzetti, L., Merla, A., Pizzella, V., Chiarelli, A.M., 2019. Deep Convolutional Neural Networks for Feature-Less Automatic Classification of Independent Components in Multi-Channel Electrophysiological Brain Recordings. IEEE Trans. Biomed. Eng. 66, 2372–2380. doi:10.1109/TBME.2018.2889512.

Crunelli, V., Lőrincz, M.L., McCafferty, C., Lambert, R.C., Leresche, N., Di Giovanni, G., David, F., 2020. Clinical and experimental insight into pathophysiology, comorbidity and therapy of absence seizures. Brain: J. Neurol. 143, 2341–2368. doi:10.1093/brain/awaa072.

ctfmeg, 2020. 000Z. CTF MEG, Canada. https://www.ctf.com/products (accessed 29 October 2020.169Z).

Davenport, E.M., Whitlow, C.T., Urban, J.E., Espeland, M.A., Jung, Y., Rosenbaum, D.A., Gioia, G.A., Powers, A.K., Stitzel, J.D., Maldjian, J.A., 2014. Abnormal White Matter Integrity Related to Head Impact Exposure in a Season of High School Varsity Football. J. Neurotrauma 31, 1617–1624.

Dekhil, O., Hajjdiab, H., Ayinde, B., Shalaby, A., Switala, A., Sosnin, D., Elshamekh, A., Ghazal, M., Keynton, R., Barnes, G., El-Baz, A., 2018. Using resting state functional MRI to build a personalized autism diagnosis system, 1381–1385. doi:10.1109/ISBI.2018.8363829.

Duan, F., Phothisonothai, M., Kikuchi, M., Yoshimura, Y., Minabe, Y., Watanabe, K., Aihara, K., 2013. Boosting specificity of MEG artifact removal by weighted support vector machine. IEEE Eng. Med. Biol. Soc. 2013, 6039–6042.

Pedregosa, Fabian, Varoquaux, Gaël, Gramfort, Alexandre, Michel, Vincent, Thirion, Bertrand, Grisel, Olivier, Blondel, Mathieu, Prettenhofer, Peter, Weiss, Ron, Dubourg, Vincent, Vanderplas, Jake, Passos, Alexandre, Cournapeau, David, Brucher, Matthieu, Perrot, Matthieu, Duchesnay, Édouard, 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Fatima, Z., Quraan, M.A., Kovacevic, N., McIntosh, A.R., 2013. ICA-based artifact correction improves spatial localization of adaptive spatial filters in MEG. Neuroimage 78, 284–294.

Garg, P., Davenport, E., Murugesan, G., Wagner, B., Whitlow, C., Maldjian, J., Montillo, A., 2017a. Automatic 1D convolutional neural network-based detection of artifacts in MEG acquired without electrooculography or electrocardiography. 2017 Int. Workshop Pattern Recognit. Neuroimag. (PRNI) 1–4. doi:10.1109/PRNI.2017.7981506.

Garg, P., Davenport, E., Murugesan, G., Wagner, B., Whitlow, C., Maldjian, J., Montillo, A., 2017b. Using Convolutional Neural networks to automatically detect eye-blink artifacts in magnetoencephalography without resorting to electrooculography. Med. Image Comput. Comput.-Assisted Interv. (MICCA) 10435, 374–381. doi:10.1007/978-3-319-66179-7_43.

Giorgio, A., Stromillo, M.L., Leucio, A.de, Rossi, F., Brandes, I., Hakiki, B., Portaccio, E., Amato, M.P., Stefano, N.de, 2015. Appraisal of brain connectivity in radiologically isolated syndrome by modeling imaging measures. J. Neurosci. 35, 550–558. doi:10.1523/JNEUROSCI.2557-14.2015.

Gonzalez-Moreno, A., Aurtenetxe, S., Lopez-Garcia, M.-E., del Pozo, F., Maestu, F., Nevado, A., 2014. Signal-to-noise ratio of the MEG signal after preprocessing. J. Neurosci. Methods 222, 56–61. doi:10.1016/j.jneumeth.2013.10.019.

Gross, J., Baillet, S., Barnes, G.R., Henson, R.N., Hillebrand, A., Jensen, O., Jerbi, K., Litvak, V., Maess, B., Oostenveld, R., Parkkonen, L., Taylor, J.R., van Wassenhove, V., Wibral, M., Schoffelen, J.-.M., 2013. Good practice for conducting and reporting MEG research. Neuroimage 65, 349–363.

Guillon, J., Attal, Y., Colliot, O., La Corte, V., Dubois, B., Schwartz, D., Chavez, M., Vico Fallani, F.de, 2017. Loss of brain inter-frequency hubs in Alzheimer's disease. Sci. Rep. 7, 10879. doi:10.1038/s41598-017-07846-w.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., Del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585, 357–362. doi:10.1038/s41586-020-2649-2.

Hasasneh, A., Kampel, N., Sripad, P., Shah, N.J., Dammers, J., 2018. Deep Learning Approach for Automatic Classification of Ocular and Cardiac Artifacts in MEG Data. J. Eng. 1–10.

Heuvel, M., Hulshoff Pol, H., 2010. Exploring the brain network: a review on resting-state fMRI functional connectivity. Eur. Neuropsychopharmacol. J. Eur. Coll. Neuropsychopharmacol. 20, 519–534. doi:10.1016/j.euroneuro.2010.03.008.

James Bergstra, Y.B., 2012. Random Search for Hyper-Parameter Optimization, pp. 281–305.

Jas, M., Engemann, D.A., Bekhti, Y., Raimondo, F., Gramfort, A., 2017. Autoreject: automated artifact rejection for MEG and EEG data. Neuroimage 159, 417–429. doi:10.1016/j.neuroimage.2017.06.030.

Barik, Kasturi, Watanabe, Katsumi, Bhattacharya, Joydeep, Saha, Goutam, 2020. Classification of Autism in Young Children by Phase Angle Clustering in Magnetoencephalogram Signals. IEEE National Conference on Communications.

Kingma, D.P., Ba, J., 2014. Adam: A Method For Stochastic Optimization.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. ImageNet classification with deep convolutional neural networks. Commun. ACM 60, 84–90.

Larson-Prior, L.J., Oostenveld, R., Della Penna, S., Michalareas, G., Prior, F., Babajani-Feremi, A., Schoffelen, J.-.M., Marzetti, L., Pasquale, F.de, Di Pompeo, F., Stout, J., Woolrich, M., Luo, Q., Bucholz, R., Fries, P., Pizzella, V., Romani, G.L., Corbetta, M., Snyder, A.Z., 2013. Adding dynamics to the Human Connectome Project with MEG. Neuroimage 80, 190–201. doi:10.1016/j.neuroimage.2013.05.056.

Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J.E., Stoica, I., 2018. Tune: A Research Platform for Distributed Model Selection and Training https://arxiv.org/pdf/1807.05118.

Abadi, Martín, Agarwal, Ashish, Barham, Paul, Brevdo, Eugene, Chen, Zhifeng, Citro, Craig, Corrado, Greg S., Davis, Andy, Dean, Jeffrey, Devin, Matthieu, Ghemawat, Sanjay, Goodfellow, Ian, Harp, Andrew, Irving, Geoffrey, Isard, Michael, Jia, Yangqing, Jozefowicz, Rafal, Kaiser, Lukasz, Kudlur, Manjunath, Levenberg, Josh, Mané, Dandelion, Monga, Rajat, Moore, Sherry, Murray, Derek, Olah, Chris, Schuster, Mike, Shlens, Jonathon, Steiner, Benoit, Sutskever, Ilya, Talwar, Kunal, Tucker, Paul, Vanhoucke, Vincent, Vasudevan, Vijay, Viégas, Fernanda, Vinyals, Oriol, Warden, Pete, Wattenberg, Martin, Wicke, Martin, Yu, Yuan, Zheng, Xiaoqiang, 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems https://www.tensorflow.org/.

McKinney, W., 2010. Data Structures for Statistical Computing in Python. In: Proceedings of the 9th Python in Science Conference. Python in Science Conference, Austin, Texas, pp. 56–61 June 28 - July 3 2010,. SciPy.

Monge, J., Gómez, C., Poza, J., Fernández, A., Quintero, J., Hornero, R., 2015. MEG analysis of neural dynamics in attention-deficit/hyperactivity disorder with fuzzy entropy. Med. Eng. Phys. 37, 416–423. doi:10.1016/j.medengphy.2015.02.006.

Muthukumaraswamy, S.D., 2013. High-frequency brain activity and muscle artifacts in MEG/EEG: a review and recommendations. Front. Hum. Neurosci. 7, 138.

Nakamura, A., Cuesta, P., Fernández, A., Arahata, Y., Iwata, K., Kuratsubo, I., Bundo, M., Hattori, H., Sakurai, T., Fukuda, K., Washimi, Y., Endo, H., Takeda, A., Diers, K., Bajo, R., Maestú, F., Ito, K., Kato, T., 2018. Electromagnetic signatures of the preclinical and prodromal stages of Alzheimer's disease. Brain: J. Neurol. 141, 1470–1485. doi:10.1093/brain/awy044.

Niso, G., Rogers, C., Moreau, J.T., Chen, L.-.Y., Madjar, C., Das, S., Bock, E., Tadel, F., Evans, A.C., Jolicoeur, P., Baillet, S., 2016. OMEGA: the Open MEG Archive. Neuroimage 124, 1182–1187.

Olde Dubbelink, K., Hilllebrand, A., Twisk, J., Deijen, J., Stoffers, D., Schmand, B., Stam, C., Berendse, H., 2014. Predicting dementia in Parkinson disease by combining neurophysiologic and cognitive markers. Neurology 82, 263–270.

Resting-State fMRI Templates – SCANlab. 2020. Resting-State fMRI Templates – SCANlab. https://brainnexus.com/resting-state-fmri-templates/ (accessed 9 November 2020).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. Int. J. Comput. Vis. 115, 211–252.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad–CAM: visual Explanations from Deep Networks via Gradient-Based Localization. In: IEEE International Conference on Computer Vision, pp. 618–626.

Simonyan, K., Zisserman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. International Conference on Learning Representations.

Smitha, S., Fox, P., Miller, K., Glahn, D., Fox, M., Mackay, C., Filippini, N., Watkins, K., Toro, R., Laird, A., Beckmann, C., 2009. Correspondence of the brain's functional architecture. Proc. Natl Acad. Sci. 106, 13040–13045.

Tadel, F., Baillet, S., Mosher, J.C., Pantazis, D., Leahy, R.M., 2011. Brainstorm: a user-friendly application for MEG/EEG analysis. Comput. Intell. Neurosci. 2011.

Tadel, F., Bock, E., Baillet, S., 2020. 000Z. Brainstorm Documentation: Visual exploration. https://neuroimage.usc.edu/brainstorm/Tutorials/ExploreRecordings (accessed 29 October 2020.332Z).

Tutorials/Epilepsy - Brainstorm 2021. Tutorials/Epilepsy - Brainstorm. https://neuroimage.usc.edu/brainstorm/Tutorials/Epilepsy#Artifact_cleaning_with_ICA (accessed 13 April 2021).

van Dyck, D., Coquelet, N., Deconinck, N., Aeby, A., Baijot, S., Goldman, S., Urbain, C., Trotta, N., Wens, V., Tiège, X.de, 2020. MEG and high-density EEG

resting-state networks mapping in children. Clin. Neurophysiol. 131, 2713–2715. doi:10.1016/j.clinph.2020.09.003.

van Essen, D.C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T.E.J., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S.W., Della Penna, S., Feinberg, D., Glasser, M.F., Harel, N., Heath, A.C., Larson-Prior, L., Marcus, D., Michalareas, G., Moeller, S., Oostenveld, R., Petersen, S.E., Prior, F., Schlaggar, B.L., Smith, S.M., Snyder, A.Z., Xu, J., Yacoub, E., 2012. The Human Connectome Project: a data acquisition perspective. Neuroimage 62, 2222–2231. doi:10.1016/j.neuroimage.2012.02.018.

van Rossum, G., Drake Jr, F.L., 1995. Python Reference manual. Centrum voor Wiskunde en Informatica Amsterdam.

Wang, Q., Tian, S., Tang, H., Liu, X., Yan, R., Hua, L., Shi, J., Chen, Y., Zhu, R., Lu, Q., Yao, Z., 2019. Identification of major depressive disorder and prediction of treatment response using functional connectivity between the prefrontal cortices and subgenual anterior cingulate: a real-world study. J. Affect. Disord. 252, 365–372. doi:10.1016/j.jad.2019.04.046.

Zikov, T., Bibian, S., Dumont, G.A., Huzmezan, M., Ries, C.R., 2002. A wavelet based de-noising technique for ocular artifact correction of the electroencephalogram, 98–105. In: Proceedings of Engineering in Medicine and Biology Society meeting doi:10.1109/IEMBS.2002.1134407.