Journal of Neuroscience Methods 364 (2021) 109367

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

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

Convolutional neural networks for decoding electroencephalography responses and visualizing trial by trial changes in discriminant features

Florence M. Aellen [a,*], Pinar Göktepe-Kavis [a], Stefanos Apostolopoulos [b], Athina Tzovara [a,c,d,*]

[a] Institute of Computer Science, University of Bern, Switzerland
[b] RetinAI Medical AG, Switzerland
[c] Sleep Wake Epilepsy Center - NeuroTec, Department of Neurology, Inselspital, Bern University Hospital, University of Bern, Switzerland
[d] Helen Wills Neuroscience Institute, University of California, Berkeley, United States

ARTICLE INFO

Keywords:
Electroencephalography
Deep learning
Convolutional neural networks
Multivariate pattern analysis
Classification
Feature extraction

ABSTRACT

Background: Deep learning has revolutionized the field of computer vision, where convolutional neural networks (CNNs) extract complex patterns of information from large datasets. The use of deep networks in neuroscience is mainly focused to neuroimaging or brain computer interface -BCI- applications. In electroencephalography (EEG) research, multivariate pattern analysis (MVPA) mainly relies on linear algorithms, which require a homogeneous dataset and assume that discriminant features appear at consistent latencies and electrodes across trials. However, neural responses may shift in time or space during an experiment, resulting in under-estimation of discriminant features. Here, we aimed at using CNNs to classify EEG responses to external stimuli, by taking advantage of time- and space- unlocked neural activity, and at examining how discriminant features change over the course of an experiment, on a trial by trial basis.

New method: We present a novel pipeline, consisting of data augmentation, CNN training, and feature visualization techniques, fine-tuned for MVPA on EEG data.

Results: Our pipeline provides high classification performance and generalizes to new datasets. Additionally, we show that the features identified by the CNN for classification are electrophysiologically interpretable and can be reconstructed at the single-trial level to study trial-by-trial evolution of class-specific discriminant activity.

Comparison with existing techniques: The developed pipeline was compared to commonly used MVPA algorithms like logistic regression and support vector machines, as well as to shallow and deep convolutional neural networks. Our approach yielded significantly higher classification performance than existing MVPA techniques ($p = 0.006$) and comparable results to other CNNs for EEG data.

Conclusion: In summary, we present a novel deep learning pipeline for MVPA of EEG data, that can extract trial-by-trial discriminative activity in a data-driven way.

1. Introduction

Multivariate pattern analysis (MVPA) is commonly used in the field of neuroscience to extract discriminative patterns of neural responses to external stimuli (Haynes and Rees, 2006). Although initially developed for functional magnetic resonance imaging (fMRI), MVPA techniques have been adapted for the field of magneto- and electro-encephalography (M/EEG) (Grootswagers et al., 2017). These are most commonly based on linear classifiers, which are applied on sensor-level topographic data, either aggregated across time (Tzovara et al., 2012) or on a time-point by time-point basis (King and Dehaene, 2014). This latter approach is most commonly implemented by training and testing one classifier at a given time-point within a trial (Castegnetti et al., 2020; Demarchi et al., 2019) and identifying time-points for which classification is above chance levels. However, this approach suffers from several drawbacks, as it only allows detecting a fixed time-period of discriminant activity for all experimental conditions and trials. Most MVPA approaches are based on single-trial information, and are thought to be more sensitive than 'classical' event-related potential (ERP) analyses. However, training and testing a classifier at single time-points makes the assumption that discriminant information appears at the same latency and electrode locations across trials, in a time- and space- locked way.

In the past few years, the field of computer vision has gained a

* Corresponding authors at: University of Bern, Institute for Computer Science, Neubrückstrasse 10, CH-3012 Bern, Switzerland
E-mail addresses: florence.aellen@inf.unibe.ch (F.M. Aellen), athina.tzovara@inf.unibe.ch (A. Tzovara).

https://doi.org/10.1016/j.jneumeth.2021.109367
Received 28 January 2021; Received in revised form 15 September 2021; Accepted 17 September 2021
Available online 23 September 2021

tremendous momentum with the introduction of deep learning algorithms (Goodfellow et al., 2016). Deep neural networks, typically relying on convolutional operations (convolutional neural networks -CNNs-) are commonly used to classify different types of images, ranging from everyday objects (He et al., 2016), to challenging medical images (Suzuki, 2017). Because the kernels of the convolutional layers share weights for the whole image, CNNs have the advantage that they are able to detect space unlocked patterns, a property called translational equivariance (Goodfellow et al., 2016, section 9.2). The position of the discriminant pattern across observations is therefore irrelevant, which often results in CNNs outperforming 'traditional' machine learning algorithms.

CNNs have been increasingly applied to new fields. In the field of EEG research, deep learning algorithms have been introduced for clinical applications such as detection of epileptic seizures (Cho and Jang, 2020; Burrello et al., 2020), automating sleep scoring (Fiorillo et al., 2019), or predicting outcome of coma patients (Jonas et al., 2019). Applications in basic research mainly focus on brain computer interfaces, oftentimes on paradigms based on motor imagery (Schirrmeister et al. 2017; Zhang et al., 2019), and (Roy et al., 2019) for a review on deep learning and EEG. Apart from BCI applications, deep learning techniques for basic EEG research such as MVPA have been introduced but are not widely used in basic research yet. Deep neural networks predominantly profit from extracting features from minimally processed data, yet several algorithms for EEG are based on hand crafted features, such as classification of time frequency (Ghosh et al., 2018; An et al., 2014), or frequency transforms of EEG data within different frequency bands (Kuanar et al., 2018; Bashivan et al., 2016; Tan et al., 2018) or differential entropy (Wang et al., 2018). While these hand crafted features often times have a physiological meaning, such as representing the energy spectrum in a given frequency band, they require making strong assumptions about the underlying task, are not easily translatable across experimental setups and might not fully exploit the features which are most discriminant across experimental conditions. Other deep learning algorithms for EEG use the raw EEG or a minimally processed signal (Schirrmeister et al., 2017; Lawhern et al., 2018; Tang et al., 2016; Nurse et al., 2016; Hajinoroozi et al., 2017), allowing the network to fine-tune its parameters and identify the most discriminant features in the data, by maximizing separability between conditions of interest. The learnt features here usually have not a physiological meaning and their interpretation is oftentimes very complex.

One important aspect common to all MVPA approaches, for basic research and also for clinical or BCI applications, is that of obtaining interpretable features that have an electrophysiological meaning (Haufe et al., 2014). Existing techniques for interpreting results of decoding approaches for M/EEG data provide information about sensor-locked activity, such as the weights or activations of single electrodes or sensors (Haufe et al., 2014). These techniques provide information about the sensors that mostly contribute to an accurate classification, but are not informative about which of the experimental conditions are driving this classification. Other techniques for feature extraction consist of separating the EEG signal in subcomponents that can be then used to visualize condition-specific patterns, like common spatial patterns (CSP) (Koles et al.,1990), but have the limitation of poor temporal resolution, and of poor generalization over multiple participants (Lotte, 2014).

In the case of CNNs, feature interpretability is its own subfield of research. One possibility for interpreting features is visualizing which dimensions of the input data are contributing to the final prediction of the network or what information the weights of the trained kernels contain (Zeiler and Fergus, 2013). This approach has the drawback that the extracted features are not trial nor condition specific. By contrast, gradient based methods, such as saliency maps, can detect discriminant patterns of activity in each individual data sample (Simonyan et al., 2013).

Here, we introduce a novel approach for decoding EEG responses to external stimuli based on CNNs. We present an MVPA pipeline, relying on a deep CNN that extracts time- and space- unlocked patterns of EEG activity; can be generalized to different datasets with minimal assumptions; and has interpretable features in terms of spatio-temporal clusters that drive an accurate classification. We explore this pipeline using two different datasets: first, a dataset consisting of EEG responses to repeated (pure tones) and novel (naturalistic) sounds, with clear and sustained differences in EEG responses. Second, we used a more challenging dataset, consisting of EEG responses to repeated and novel images, whose presentation was mixed across participants, resulting in more subtle condition differences. Our goal is to use this pipeline in order to extract in a data-driven way trial by trial spatio-temporal patterns of discriminant electrophysiological responses, and explore how these change over the course of an experiment.

2. Materials and methods

2.1. Data

We used two different EEG datasets to (a) build our pipeline, and (b) evaluate whether it generalizes across experimental settings. The first dataset (termed 'Auditory') was used to develop the presented algorithm and fine tune its individual steps. The second dataset (termed 'Visual') was in turn used to examine whether the developed pipeline can also be used on new data and experimental conditions. Both datasets are openly available (Cavanagh et al., 2018; van Peer et al., 2017).

2.1.1. Auditory dataset

The first dataset was an auditory oddball paradigm, consisting of repeated presentations of *Standard*, *Target* and *Novel* sounds. The *Standard* and *Target* sounds were sinusoidals at different frequencies, while the *Novel* ones were naturalistic sounds, varying with each presentation. Here, we considered the EEG data of 17 participants from the control group of this dataset, disregarding participants with persistent artifacts or noise in their recordings. For simplicity, we focused on a 2-class classification problem, and considered trials were participants were presented with either a *Standard* or *Novel* sound (Fig. 1 a and b for mean responses across participants, and Fig. S1 for a topographic representation). The data was recorded with 64 electrodes in a standard 10/20 configuration at a sampling frequency of 500 Hz, initially referenced to the CPz electrode. Four temporal electrodes were removed, as in the original publication of the data (Cavanagh et al., 2018), and the remaining electrodes were re-referenced to a common average reference. Additionally eye blinks were removed with Independent Component Analysis (ICA) and single trials were extracted on a time window of 0.6 s (−0.1 to 0.5 s relative to stimuli onset). We additionally filtered the data between 0.1 and 20 Hz. This first dataset contained a mean of 129.5 ± 2.7 *Standard* (mean ± standard error reported here and in the following) and 28.5 ± 1 *Novel* trials per participant.

2.1.2. Visual dataset

The second dataset was a visual oddball, consisting of a repeated presentation of different sets of images. Similar as in the auditory dataset, we considered two classes of *Familiar* and *Novel* images (Fig. 1 c and d for mean responses across participants, and Figure S1 for a topographic representation). We extracted data from 20 participants in total, disregarding participants with prominent artifacts or noise in their recordings. EEG data were recorded at 512 Hz (later down-sampled to 256 Hz) with 64 electrodes in a standard 10/20 montage, referenced to an active common mode sense (CMS). EEG data were filtered between 0.1 and 30 Hz and re-referenced to a common average reference. Eye blinks and movement were removed according to Gratton et al. (1983). Single trials were extracted on a time window of 1.5 s (0–1.5 s relative to stimuli onset). For the visual dataset we did not include any baseline, as the data that were publicly released were already epoched, without any baseline (van Peer et al., 2017). We additionally inspected single trials visually for artifacts. Noisy trials containing eye blinks or muscle activity

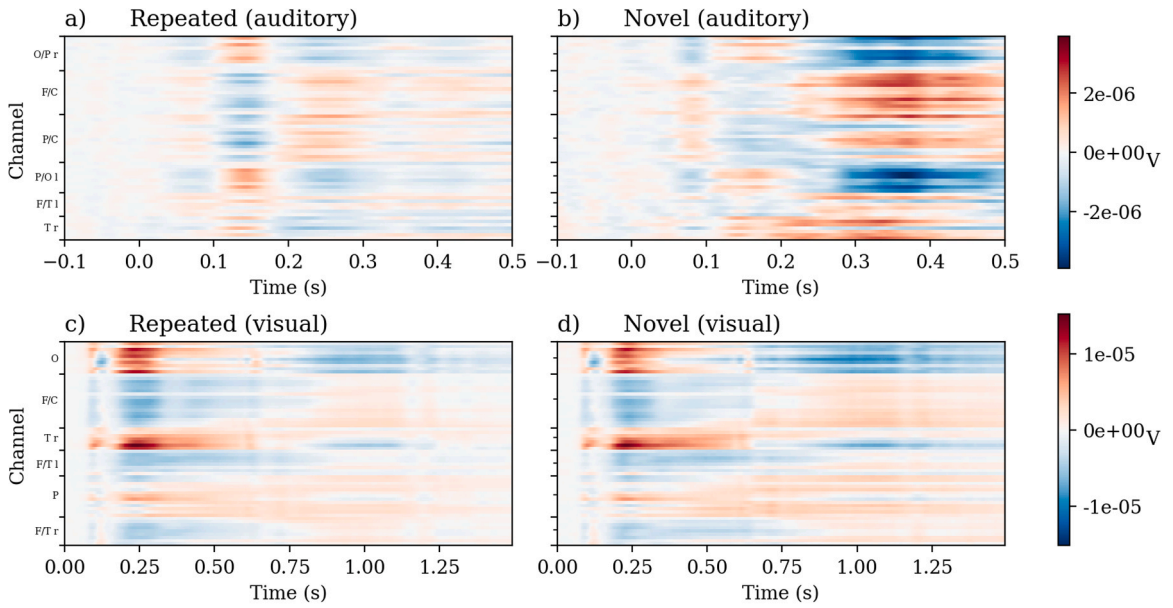

Fig. 1. Mean evoked responses for the auditory (top) and visual (bottom) dataset, represented as time by electrodes. Panels a and c show the mean of all *Repeated* trials and panels b and d the mean of all *Novel* trials. The y-axis displays the recorded EEG channels, grouped in regions of interest for illustration purposes.

were rejected. This resulted in 506 ± 140 *Familiar* trials and 146 ± 40 *Novel* trials per participant.

In the following, for reasons of consistency, we refer to the two classes of both datasets as *Repeated* (replacing *Standard* and *Familiar* from the first and second dataset respectively) and *Novel*. We represented the EEG data as a 2D signal throughout most of the rest of the paper, where the first dimension were the channels and the second the time (Fig. 1).

2.1.3. Train, validation and test sets

Each dataset (Auditory and Visual) was split into a train, validation and test set in a 10-fold procedure. We used these splits to train the neural network 10 times, in a cross validation way. The validation set was used for optimizing the network's hyper-parameters and identifying the best fold. The test set was left aside until the very end, and was never used for tuning the neural network or its hyper-parameters. The test set was only used to evaluate, in an unbiased way, the network's performance. The available data were split into 81% train, 9% validation and 10% test trials. For the auditory dataset this resulted in 2176 trials for the train, 242 for the validation and 269 for the test dataset. As for the visual dataset there were 10570 trials in the train, 1306 in the validation and 1175 in the test set.

2.2. Network architecture

We built our MVPA pipeline around a residual neural network with 50 layers (ResNet50) (He et al., 2016) (Fig. 2, red box on the left side). This network consists of a convolutional layer, batch normalization, ReLu activation and a MaxPooling layer, followed by four segments of 3/4/6/4 convolutional blocks each. Each convolutional block has three convolutional layers followed by a batch normalization and ReLu activation layer. After the last batch normalization layer, the original input to the convolutional block is added to the output from the batch normalization layer. This residual skip connection is the main novelty of the ResNet architectures compared to most convolutional neural networks. The skip connections allow for deeper networks, which can extract more complex structures from the input data. For the first convolutional block of all the four segments there is an additional convolutional and batch normalization layer on the skip connection due to otherwise mismatched dimensions. All further technical information, such as kernel sizes and padding information can be found in (He et al., 2016). In addition to the standard architecture of ResNet50, we additionally included a fully connected layer with 128 nodes and a dropout layer (with 50% probability), to further restrict overfitting. EEG trials, represented as (Channels) × (Time) were given as input to the network. The network's output was a probability value per trial, ranging between zero and one, describing the probability of this trial to belong to each of the two classes (the *Repeated* class was assigned the label 0).

2.3. Data augmentation

Data augmentation techniques are commonly used in the field of computer vision, to artificially increase the size of an existing dataset and avoid overfitting (see Shorten and Khoshgoftaar, 2019) for an overview of data augmentations used in computer vision). These techniques essentially distort parts of the input data in a minor but meaningful way before training a network. For example, in the field of computer vision, commonly used data augmentation techniques consist of flipping an input image horizontally, which is ecologically valid, as it is possible to observe object rotations in nature. In the case of EEG data, which are time series, flipping the time dimension would not make sense.

Here, taking into account the nature of EEG, we augmented the available trials in three different ways: (a) time shifts; (b) Gaussian noise and (c) sub-averaging single trials (Fig. 3). First, in order to account for inter-individual differences in the timing of EEG responses, the available single trials were shifted in time with a random interval of up to 5 time-points in either the positive or negative direction. Second, to account for different levels of noise across participants, we additionally augmented the data by adding Gaussian noise, with a mean of zero and a random standard deviation of 0.1, 0.2 or 0.3 per trial. Third, considering the noisy nature of single-trial scalp EEG responses, we averaged the input data over multiple trials. More specifically, per trial we chose a random number n_k ($k \in (1, b)$) (where b is the batch size) from a triangular distribution (centered around 1) between 1 and 21 (which corresponds to 1/3 of the trials with the same labels in the current batch). For each trial we then chose $n_k - 1$ samples from that batch with the same labels and took the mean over the original and the additional trials. This last technique of averaging single trials is commonly used in classification of EEG responses, in order to improve signal-to-noise ratio (Tzovara et al.,

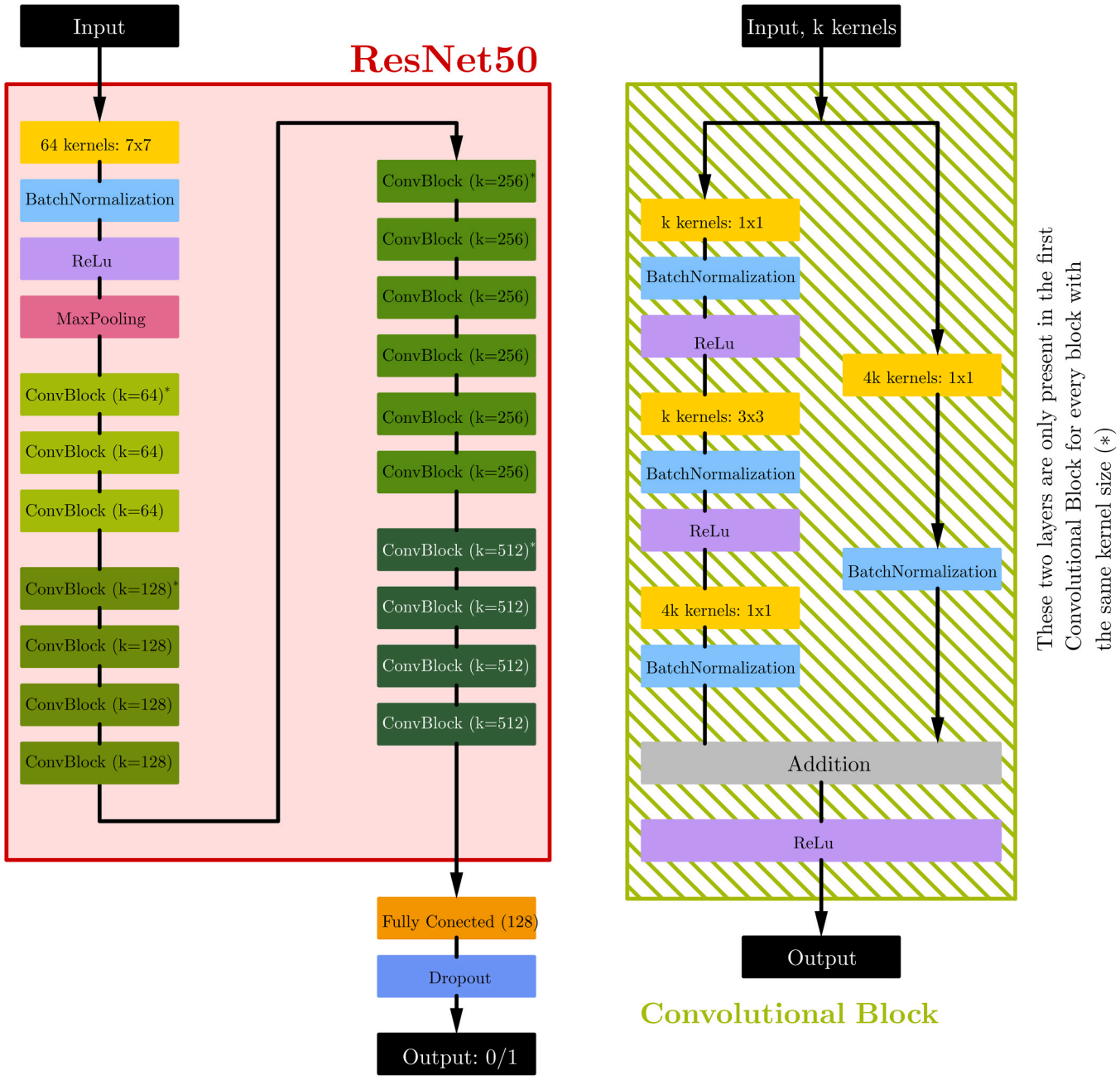

Fig. 2. Schematic representation of the architecture of the neural network. The main network has four sections of each 3/4/6/4 convolutional blocks. Each convolutional block then contains three convolutional layers. We added a fully connected and dropout layer on top of the ResNet architecture. The network outputs either 0 or 1, for the labels *Repeated* or *Novel*.

2012).

2.4. Optimization and training

In the two datasets used here, the number of trials in the two classes was imbalanced, with a ratio of *Repeated* to *Novel* trials of roughly 4–1. To account for this imbalance, we over-sampled the underrepresented (*Novel*) class during training, by drawing an equal number ($b/2$) of trials from two pots containing all trials from the training set of each of the two classes.

In the training pipeline, a batch of size b of data, containing ($b/2$) trials from each condition went trough the data augmentation step before training. To account for the imbalance during validation and test, we measured the area under the Receiver Operator Characteristic curve (AUC) (Macmillan and Creelman, 2004), which consists of the true positive vs. false positive rate with respect to multiple thresholds.

The network was optimized with an Adam optimizer, using the standard parameters proposed in its original implementation (Kingma and Ba, 2014), to minimize the binary cross-entropy loss (Eq. 1) between the real labels y of the data and the network's predictions $\hat{y}$.

$$L = -(y \log(\hat{y}) + (1-y) \log(1-\hat{y})) \quad (1)$$

During training we employed early stopping, so that if the validation loss would not further decrease within ten training epochs the training would stop (Goodfellow et al., 2016) section 7.8). The network was trained for maximally 50 epochs. As a final step, we retained the network with the smallest validation loss. In the Results section 3.1 we report the mean train, validation and test AUC score, accuracy and

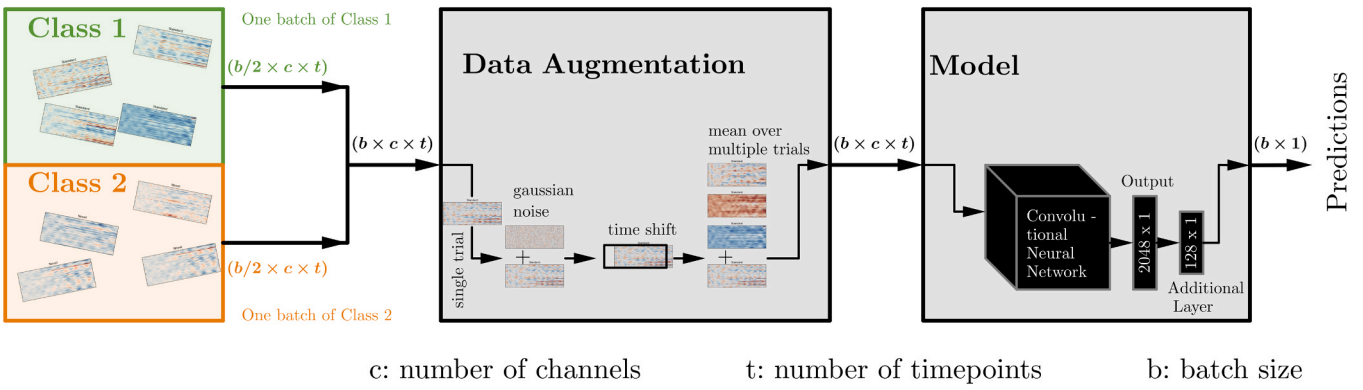

Fig. 3. Pipeline for data augmentation and training. The pipeline starts with selection of samples in the current batch (left panel), proceeds with data augmentation (central panel) and last inputs the trials into the network for training.

binary cross-entropy loss at that best epoch for all trained networks.

The network layers were initialised with imagenet weights (Fchollet, 2016) and the batch size b was chosen to be of 64 samples, because of GPU size limitations. We used the python library tensorflow (2.1.0) with cuda (10.1.168), cudnn (5.1.5) and python (3.6.8). The full training pipeline with the two classes, data augmentation and the CNN is illustrated in Fig. 3.

2.5. Estimation of network performance

To evaluate the network's performance, we computed chance levels in a data-driven way, through random permutations. We randomly shuffled the labels of the training dataset 50 times. For each random shuffle we retrained the networks in the 10-folds of the cross-validation, resulting in 500 'random' networks. Similar as for the networks trained on true labels, we retained the test scores at the epochs of smallest validation loss. We then used these 'random' networks to classify trials from the Test and Validation dataset. Each of the permutations resulted in one chance level classification performance. The actual performance of the network, trained with true labels, was compared to the distribution of AUC values obtained with random permutations with a Wilcoxon signed-rank test. Due to the heavy computational cost of training CNNs, and due to the overwhelmingly low chance-level classification results that we obtained for the first dataset when permuting the true labels, we only computed chance levels via random permutations for the first dataset. For the second dataset (Visual) we compared instead the performance of the CNN with a 'classical' MVPA approach (see section 2.8.1).

2.6. Discriminant features

To visualize features from the EEG data that mostly contribute to the network's output, we used a gradient-based technique termed saliency maps (Simonyan et al., 2013). This technique consists of back-propagating the input label of a given trial through the network, to obtain the gradient, i.e. a value per time point and electrode marking the strength of the contribution of that input point to the decision of the network. For a given input I_0, output class c and a score function $S_c(I_0)$ (here binary cross entropy loss), the gradient w of S_c with respect to an input I at the point I_0 is given as

$$w = \frac{\partial S_c}{\partial I}|_{I_0}. \tag{2}$$

The absolute value of w then gives the saliency map. To calculate saliency maps (in the following called activation maps), we used the implementation from Kotikalapudi and contributors (2017). For each fold of the 10-fold cross validation, we trained four networks, resulting in a total of 40 networks. This follows feature visualization approaches that are commonly used in the biomedical field, where physiological data are more noisy and complex compared to natural images Fauw et al. (2018). Typically, multiple networks are trained per fold and their outputs are averaged to obtain stable features that are consistently identified by all networks (Fauw et al., 2018; Mehrer et al., 2020). Here, we calculated activation maps for each of the 40 networks separately, and then averaged the mean of the obtained maps. This resulted in one activation map per input data point, which were either average ERPs for the two experimental conditions (section 3.4) or single trial ERPs (section 3.6).

2.7. Trial by trial representation of discriminant features

As activation maps can be computed at the single-trial level, they can be used to study changes in trial-by-trial neural responses throughout an experiment. Instead of considering all single-trial EEG responses, as in most ERP or MVPA analyses, with single-trial activation maps it is possible to retain the temporal order of trials and compare their evolution from the first, the second, up to the last presentation of the stimuli. Such an approach could allow to assess for instance effects of learning, where neural responses to a given stimulus change from one trial to the next as a function of presentation.

Here, to explore the potential of single-trial feature extraction, as an exemplar test case, we extracted a sequence of single-trials keeping the order of exposure of the participants to the stimuli of each experimental condition. We then averaged single-trial responses over participants for each consecutive stimulus presentation, and calculated the activation maps for each of these responses. This resulted in a sequence of activation maps, which reflect patterns of discriminant EEG activity across consecutive presentations of the experimental stimuli. To quantify changes in activation maps over the course of the experiment, at every trial and time point we summed up the values of activation maps over all electrodes (Figure S4 in the Supplemental Material), resulting in one value per trial and time-point. We then fitted a linear regression at every time point to test whether the overall discriminant EEG activity changed significantly from zero as a function of trial repetitions throughout the experiment. To correct for multiple comparisons across time, 1000 cluster-based permutations were used (Maris and Oostenveld, 2007). With this analysis we identified time points with a significant change in the activation of the network over the course of the experiment.

As a control, we performed the same analysis on the EEG data. Instead of the activation maps we used EEG activity at every time point recorded across electrodes, and tested whether there were any latencies where changes in EEG responses during the experiment were significantly different from zero as a function of trial repetitions.

2.8. Comparison with existing techniques

2.8.1. Comparison to logistic regression and support vector machines

The performance of the CNN was compared to two baseline algorithms, using exemplar MVPA techniques. For this comparison, we chose logistic regression and support vector machines (SVM with 'rbf' kernel), as they are commonly used in the field of M/EEG research (Tomioka et al., 2006; Castegnetti et al., 2020; Philiastides et al., 2010). For this comparison, we kept the same splits of train/test/validation and the same cross-validation procedure as for training and testing the CNN. To estimate the hyper-parameters of the logistic regression, we pooled all observations from the Training set together, and optimized the parameters of penalty (l1 or l2 norm) and inverse regularizer (0.01–100, with logarithmic spacing). The optimized parameters were then used to train and test one classifier for every time point, resulting in one classification score per time-point. This resulted in one time course of training and test AUC values averaged over the 10-fold cross-validation. For SVM, we use the same approach for optimizing the hyper-parameters gamma ('scale' or 'auto') and regularization parameter (0.01–100 with logarithmic spacing). To compare the performance of the logistic regression and SVM with the CNN, we retained the best performance of these two algorithms, at the point of the maximum validation AUC score, and contrasted this with the overall performance of the CNN. Same as for the CNN, chance levels for logistic regression were evaluated by shuffling the labels of the training dataset, and by retraining the classifier of each time-point 50 times. The performance of the real classifier was compared with the distribution of the performance of chance classifiers using Wilcoxon signed-rank tests and was cluster-based corrected for multiple comparisons over time (Maris and Oostenveld, 2007).

2.8.2. Comparison with other CNN architectures

We also compared the performance of the ResNet-based CNN to two other CNNs that are commonly used for decoding EEG signals (Zhang et al., 2019; Ghosh et al., 2018; Williams et al.,2020; Jonas et al., 2019). We used a *Shallow* and *Deep* CNN, first introduced in Schirrmeister et al. (2017). These consist of 2 and 5 convolutional layers, for the *Shallow* and *Deep* networks respectively, with additional batch normalization, activation, pooling and dropout layers in between. They don't have any residual skip connections and are therefore shallower than the 50 layered ResNet. For a fair comparison across all CNN architectures (ResNet, *Shallow* and *Deep* networks), we always used the same training and data augmentation pipeline as described in section 2.4.

Additionally, we evaluated the effect of some of the choices made in the training pipeline introduced here in classification performance (S.2.1, S.2.3, S.2.2). More specifically, we compared a CNN trained with filtered vs. with unfiltered data, a CNN where the underrepresented class was oversampled to a CNN where a weighted binary crossentropy loss was used, and lastly a CNN where we omitted the fully connected layer with 128 nodes before the dropout layer. Details for these control analyses can be found in the Supplemental Material.

3. Results

3.1. Training and classification performance of the CNN pipeline

First, we trained CNNs to classify EEG responses to *Repeated* vs. *Novel* stimuli, using the training, validation and test folds as described in section 2.1. For the auditory dataset, decoding performance, measured through the AUC, increased for both train and validation sets already within the first 10 epochs of training, and reached a plateau approximately from epoch 20 on (Fig. 4, panel a). At the same time, the binary cross-entropy loss decreased and reached a plateau already after the first 10 epochs of training (Fig. 4 panel c). We report an AUC score of 0.89 ± 0.04 on the train, 0.75 ± 0.04 on the validation and 0.72 ± 0.04 on the test set (see Table 1). On average across folds, these scores were reached on epoch 28.7 ± 13.7. The high classification performance in the test set suggests that the trained networks could extract discriminant features of EEG responses to *Repeated* vs. *Novel* sounds, and generalize to

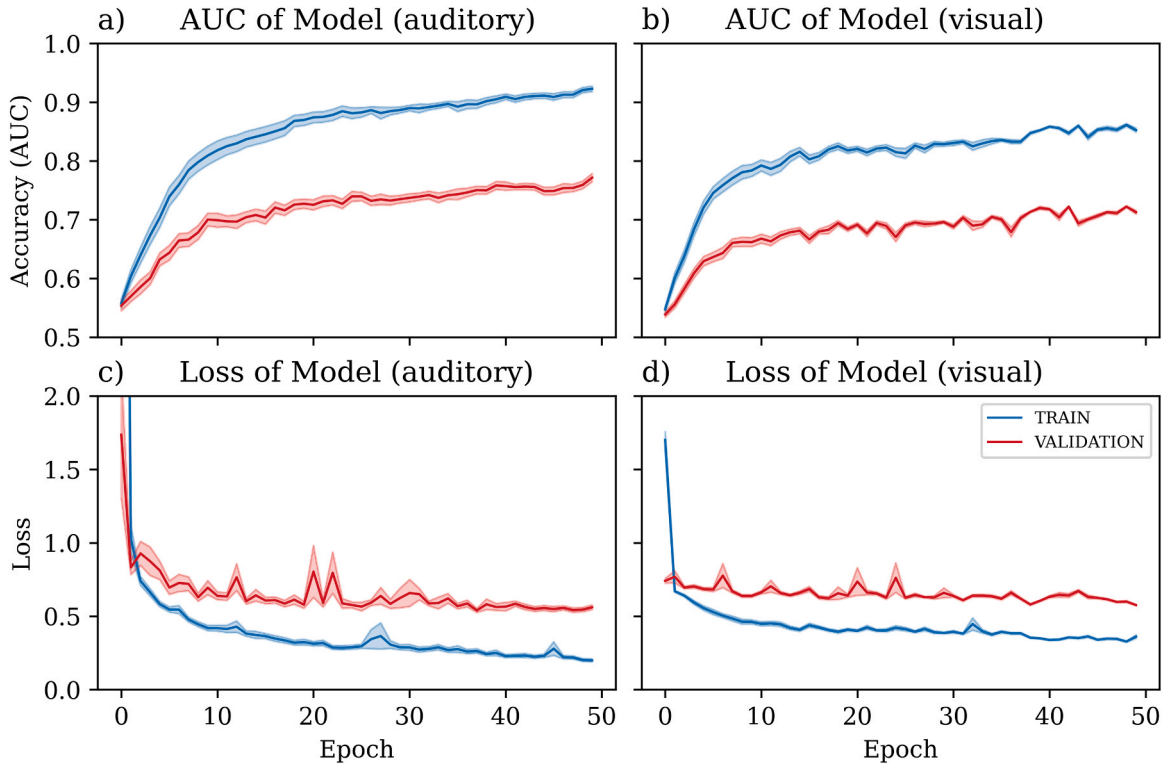

Fig. 4. Training performance of the CNN on the two datasets. In each plot, the bold line illustrates the mean AUC (panels a and b) or binary cross-entropy loss (panels c and d) over the 40 trained networks and the shaded area the standard error. Blue lines correspond to the scores of the train and red lines to the scores of the validation set, respectively. Panels a and b show the results for the auditory dataset and panels c and d for the visual dataset.

Table 1
Results of training the *Shallow*, *Deep* convolutional neural network and the ResNet for both datasets. We report the binary cross-entropy loss, the AUC score and the accuracy for the train, validation and test sets. The reported results correspond to the mean scores ± standard error at the epoch with the lowest binary cross-entropy loss on the validation set.

	Binary cross-entropy loss			AUC-score			Accuracy		
	Train	Validation	Test	Train	Validation	Test	Train	Validation	Test
Auditory dataset									
ResNet	0.26 ± 0.08	0.43 ± 0.05	1.16 ± 2.33	0.89 ± 0.04	0.75 ± 0.04	0.72 ± 0.04	0.82 ± 0.12	0.79 ± 0.11	0.77 ± 0.10
Shallow Net	0.38 ± 0.01	0.37 ± 0.03	0.44 ± 0.04	0.83 ± 0.01	0.78 ± 0.04	0.75 ± 0.03	0.86 ± 0.01	0.85 ± 0.02	0.82 ± 0.02
Deep Net	0.39 ± 0.03	0.379 ± 0.03	0.47 ± 0.04	0.83 ± 0.01	0.77 ± 0.03	0.73 ± 0.03	0.85 ± 0.01	0.85 ± 0.02	0.82 ± 0.03
Visual dataset									
ResNet	0.40 ± 0.06	0.54 ± 0.03	0.61 ± 0.06	2.82 ± 0.03	0.69 ± 0.02	0.67 ± 0.04	0.76 ± 0.04	0.75 ± 0.03	0.75 ± 0.04
Shallow Net	0.52 ± 0.01	0.48 ± 0.01	0.52 ± 0.04	0.74 ± 0.01	0.67 ± 0.03	0.68 ± 0.02	0.79 ± 0.02	0.78 ± 0.01	0.78 ± 0.02
Deep Net	0.54 ± 0.02	0.49 ± 0.01	0.58 ± 0.06	0.73 ± 0.01	0.63 ± 0.02	0.68 ± 0.01	0.76 ± 0.02	0.76 ± 0.02	0.76 ± 0.02

new, previously unseen trials.

As a next step, we evaluated whether the training pipeline also generalized to a different dataset and modality. The network training for the visual dataset proceeded in a very similar way as for the auditory. The AUC and the binary cross-entropy loss reached a plateau at around epoch 20 (Fig. 4, panels b and d). The maximum AUC score was 0.82 ± 0.03 on the train, 0.69 ± 0.02 on the validation and 0.67 ± 0.04 on the test set (see Table 1), and corresponded to epoch 17.1 ± 10.4. This result suggests that the training pipeline, even though developed for the auditory dataset, can be applied to a different dataset.

3.2. Comparison of the network performance to chance levels

The network's performance was contrasted to chance levels, computed by re-training the CNN on data with randomly shuffled labels (Fig. 5). During training, the AUC score in the training set slightly increased with training. However, the AUC in the validation set remained around the baseline values of 0.5 (Fig. 5, panel a). A similar tendency was observed for the binary cross-entropy loss, which decreased over the first few epochs of training, but remained around 0.69 (log(2), corresponding to theoretical chance levels) for all validation epochs (Fig. 5, panel b). Importantly, networks trained on real data achieved a consistently higher AUC compared to networks trained on data with shuffled labels, across all folds of the cross-validation (Fig. 5, panel c) (Wilcoxon signed-rank test, p = 1.2e-83).

3.3. Comparison with existing techniques

3.3.1. Logistic regression and support vector machine

To compare the results obtained with the CNN with existing techniques, we trained linear and non linear 'classical' MVPA algorithms to discriminate *Repeated* vs. *Novel* stimuli in both datasets. For every time point, we trained and tested classifiers based on logistic regression (linear classifier), and SVM with 'rbf' kernel (for a non linear classifier), which resulted in a time course of AUC values, computed on a train and test set (Fig. 6 panel a for the auditory and b for the visual datasets).

Using logistic regression in the auditory dataset, the classification score of the test set was around 0.5 during the 0.1 s before stimulus onset (Fig. 6 panel a −0.1 to 0.0 s), and increased after the stimulus onset, reaching a maximum AUC score of 0.63 ± 0.01 on the test set, across folds, at 0.323 ± 0.01 s post-stimulus onset.

In the visual dataset, there was no baseline in the available data (van Peer et al., 2017). Decoding performance started to increase around 0.1 s post-stimulus onset (Fig. 6 panel b). The maximal decoding performance on the test set was 0.62 ± 0.01, and was reached at 0.70 ± 0.10 s post stimulus onset.

Chance levels, estimated through random permutations, were on average 0.5 throughout the entire trial interval (Fig. 6, panels a and b, green lines) for both datasets. Decoding performance was significantly above chance levels from 0.032 to 0.5 s post stimulus onset for the auditory, and from 0.0 to 1.5 s post stimulus onset for the visual datasets (Fig. 6, panels a and b, marked in gray lines).

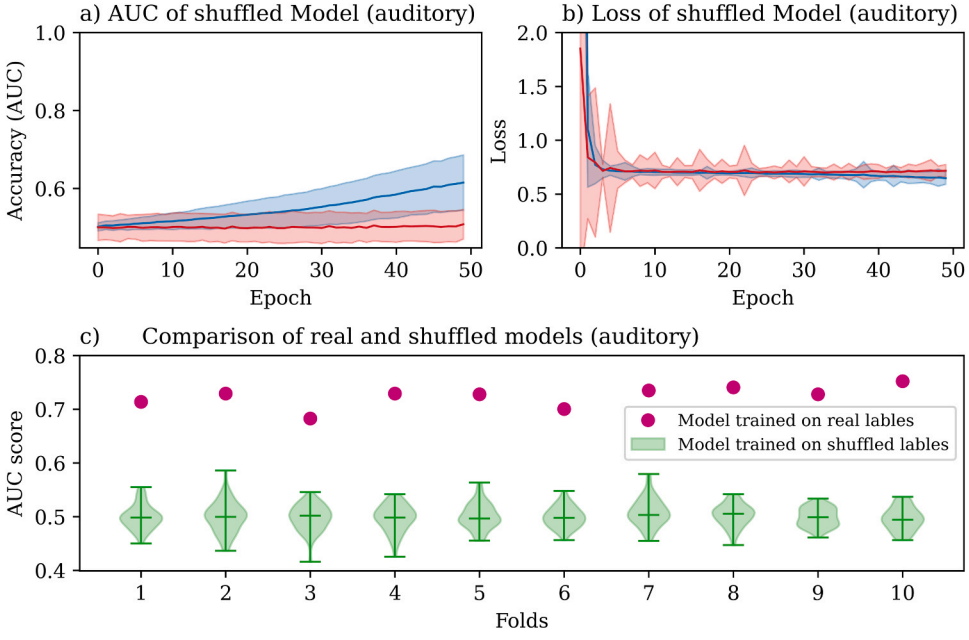

Fig. 5. Comparison of the CNN classification performance with chance level results of the shuffled CNNs on the auditory dataset. The bold lines illustrate the mean AUC score (panel a) or loss (panel b) over the networks trained with random permutations, and the shaded area the standard error. The blue line shows the scores of the train and the red line the scores of the validation set. Panel c shows the comparison of true and chance level classification performance on the test set. For each fold we show the mean performance of the real network (magenta) vs. the distribution of the performance of the 500 shuffled networks (green violin plots).

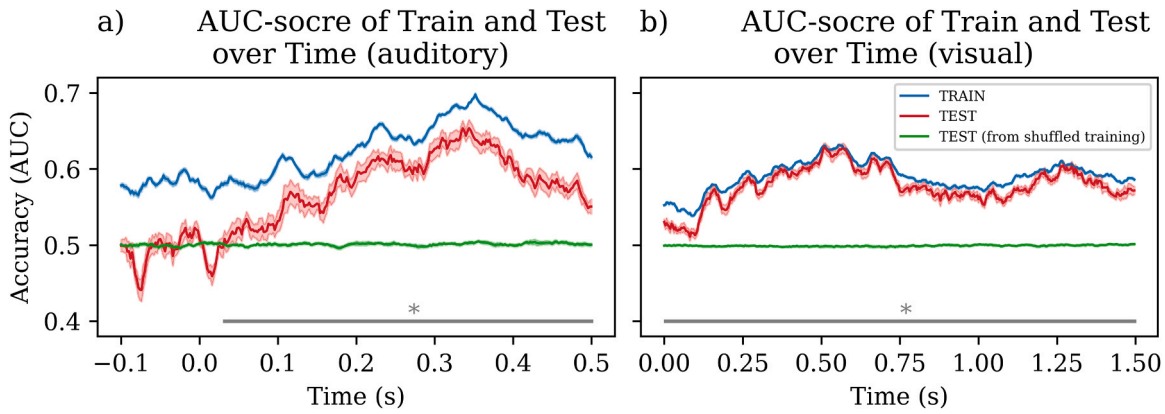

Fig. 6. Classification performance for a logistic regression classifier trained and tested at every time point separately. We report the AUC scores of the train (blue) and test set (red), averaged over 10-folds of cross-validation, and the scores of the test set, in the case where the classifier was trained on shuffled data (green). Bold lines show the mean scores over the 10-fold validation and the faded colored regions show the standard error. Horizontal gray lines show the time-periods where classification was significantly above chance levels in the two datasets.

We also trained and tested a SVM, with the same procedure as for logistic regression. The maximum AUC score reached on the test set was on average 0.58 ± 0.01 at 0.352 ± 0.02 s post-stimulus onset for the auditory and 0.60 ± 0.004 at 0.59 ± 0.06 s post-stimulus onset for the visual dataset.

Next, we contrasted the performance of the CNN with the two baseline MVPA algorithms. For this comparison, we contrasted the maximum classification performance obtained across time with logistic regression and SVM, to the overall performance obtained with the CNN. This approach is rather conservative, and might penalize the CNN. Nevertheless, in each of the 10 folds of the cross validation, the CNN provided higher classification performance than both the logistic regression and the SVM (Fig. 7). For both the auditory and visual datasets, the AUC of the CNN was significantly higher than the AUC of logistic regression (Wilcoxon signed-rank test, $p = 0.006$ for both datasets, corrected for multiple comparisons), and than the AUC of SVM (Wilcoxon signed-rank test, $p = 0.006$ for both datasets, corrected for multiple comparisons, Fig. 7).

3.3.2. Comparing different CNN architectures

Additionally, we compared the ResNet50-based pipeline that we developed to other existing CNNs that have been previously used on EEG data, including *Shallow* and *Deep* CNNs. The AUC score obtained on the test set with the *Shallow* CNN was 0.75 ± 0.03 for the auditory and 0.68 ± 0.02 for the visual dataset. The *Deep* CNN resulted in an AUC of 0.73 ± 0.03 and 0.68 ± 0.01 for the auditory and visual datasets, respectively (Table 1 and Fig. 8). There was no significant difference in AUC values for ResNet vs. *Shallow*, ResNet vs. *Deep*, or *Shallow* vs. *Deep* networks (Wilcoxon signed-rank test, $p > 0.08$, corrected for multiple comparisons). These results suggest that the developed pipeline can classify EEG data under different network architectures.

Last, we evaluated the robustness of the developed pipeline under slight modifications in the pipeline architecture or input data (Supplemental material S.2). Notably, the classification performance remained at similar levels for filtered and unfiltered data (S.2.1), or when omitting the final fully connected layer with 128 nodes before the dropout layer of the CNN (S.2.3). Oversampling of the underrepresented class yielded a higher classification performance than using a weighted binary crossentropy loss (S.2.2).

3.4. Extraction of discriminant features

After establishing that the networks can accurately classify *Repeated* from *Novel* stimuli, we next visualized the discriminant features that were driving this classification. Fig. 9 shows the activation maps of the mean EEG responses to *Repeated* and *Novel* auditory (panels a and b) and visual (panels c and d) stimuli (Fig. 1). In the representation of activation maps, stronger colors denote that a given electrode and time interval were more relevant in the network's output than lighter ones. For the auditory dataset (Fig. 9, panels a and b), almost all of the non zero activations appeared after stimulus onset. The highest activation values occurred at different latencies for each experimental condition, ranging from 0.2 to 0.3 s for the *Repeated* and from 0.3 to 0.5 s for the *Novel* trials. For the visual dataset (Fig. 9, panels c and d), the two experimental conditions (*Repeated* and *Novel*) had a more similar distribution of activations. Most of the non-zero activations for the visual dataset

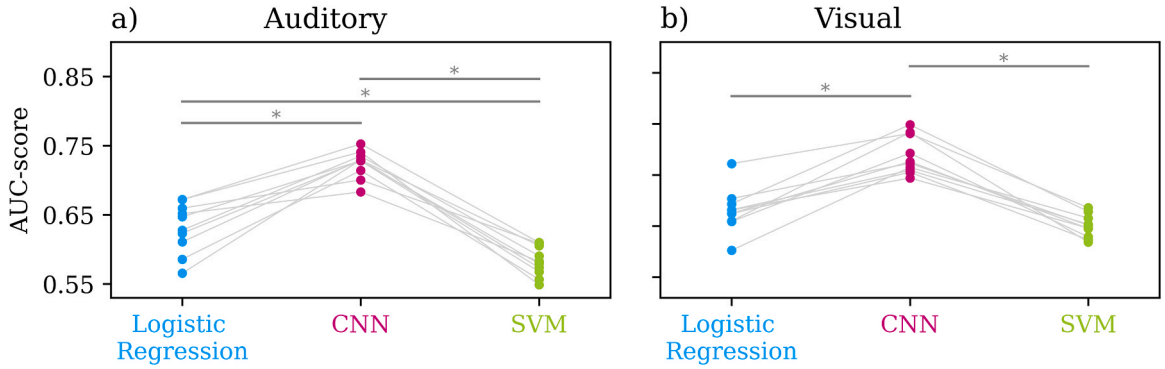

Fig. 7. Comparison of the performance of logistic regression vs. CNN vs. SVM for the auditory dataset (panel a) and comparison of logistic regression vs. CNN vs. SVM for the visual dataset (panel b). Each dot corresponds to the test sets of one of the 10 folds of the cross validation. For the CNN we report the mean AUC score over the 4 trained networks per fold.

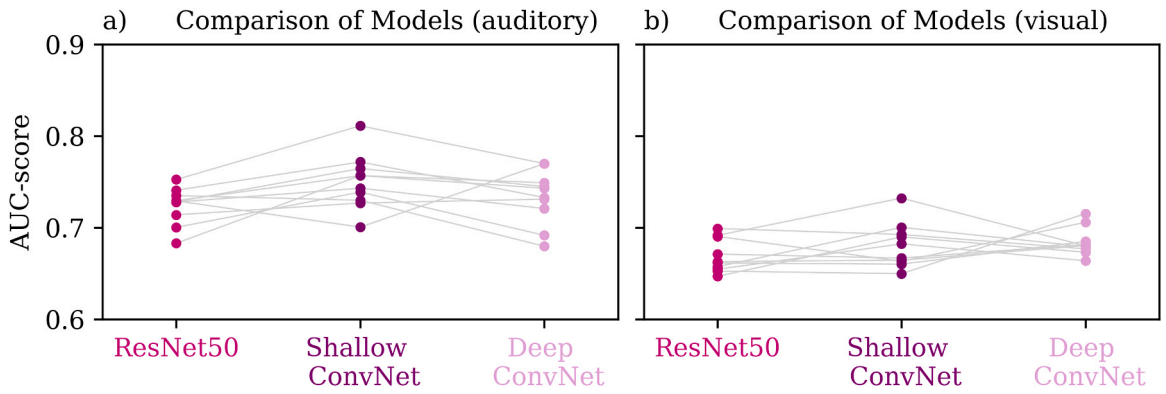

Fig. 8. Comparison of the performance of the *Shallow*, *Deep* convolutional neural networks and ResNet50. The difference in performance was not significant for all comparisons after correcting for multiple comparisons (Wilcoxon signed-rank test, $p > 0.08$, corrected for multiple comparisons).

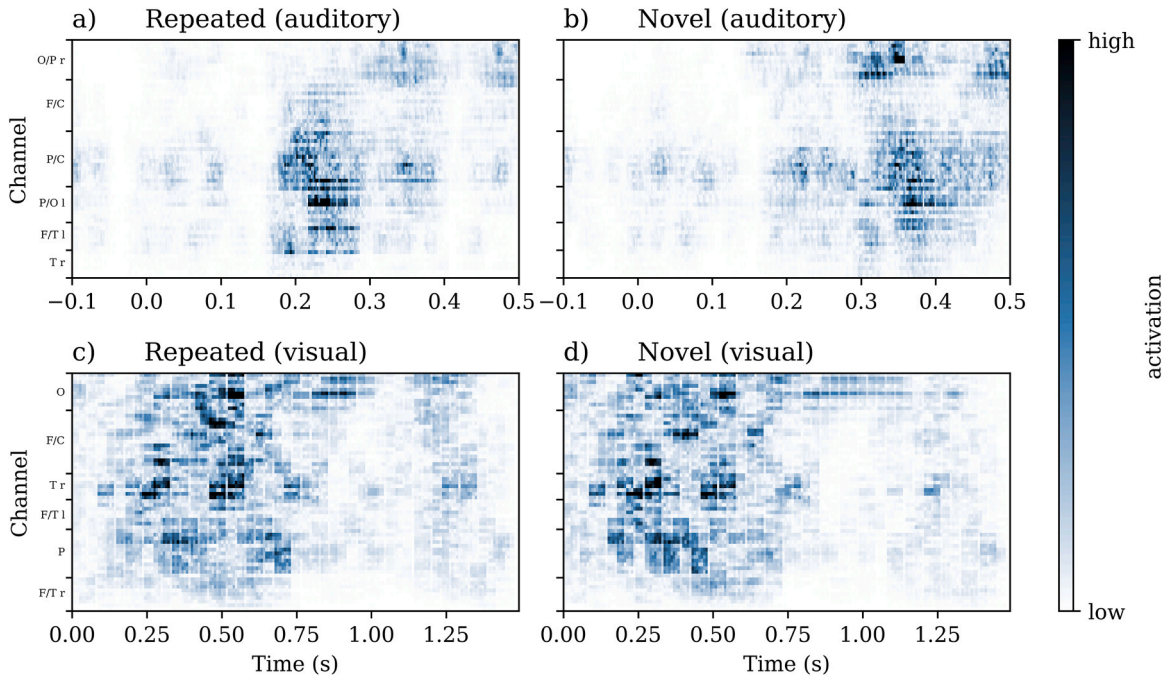

Fig. 9. The activation maps of the mean trials of the auditory (top) and visual (bottom) dataset for *Repeated* (panels a and c) and *Novel* (panels b and d) trials.

occurred between 0.25 and 0.7 s, at similar latencies for both experimental conditions.

3.5. Topographic representation of discriminant features

As an alternative representation, the activations of the CNN were additionally visualized as topographic maps, by reassigning the electrodes to their original location on the scalp. Fig. 10 shows the activation maps for each dataset and condition as a topographic map, to give a more interpretable visualization of the networks' features. For reasons of consistency, Fig. 10 provides similar latencies for both datasets. Each topographic map displayed the sum of activations in steps of 0.1 s, as described above the map (Fig. 10). This topographic representation revealed that the discriminant information for the *Repeated* auditory condition started appearing at 0.2–0.3 s post stimulus onset, mainly at fronto-occipital electrodes (Fig. 10, panel a). By contrast, for the *Novel* auditory condition, discriminant information was more strongly appearing at centro-parietal electrodes, between 0.3 and 0.4 s post-stimulus onset (Fig. 10, panel b), matching closely the actual topographic maps of the data (Fig. S1). For the visual dataset, activations were stronger at similar latencies for *Repeated* and *Novel* conditions, starting mainly after 0.2 s post-stimulus onset, and occurring predominantly at central and occipital electrodes (Fig. 10, panels c and d), closely following the topographic maps of the average data (Fig. S1).

3.6. Trial by trial changes in discriminant features

As the features of activation maps can be computed at the single-trial level, we performed an exploratory analysis, evaluating trial by trial changes in the activation maps throughout the experiment. Fig. 11 illustrates the time-course of a linear regression analysis, quantifying trial by trial changes on the activation maps (panel a). The slope of the linear regression was significantly different from zero from 0.106 to 0.272 s for the *Repeated* condition ($p < 0.05$, corrected with cluster-based permutations) (Fig. 11, orange horizontal line). For the *Novel* condition, there was no period of significant change in the slope of the linear regression throughout the experiment (Fig. 11, green line).

The same analysis was performed on the EEG data (section 2.7). For the EEG data, the slope of the linear regression was close to zero for both conditions throughout the entire temporal interval (Fig. 11 panel b), and

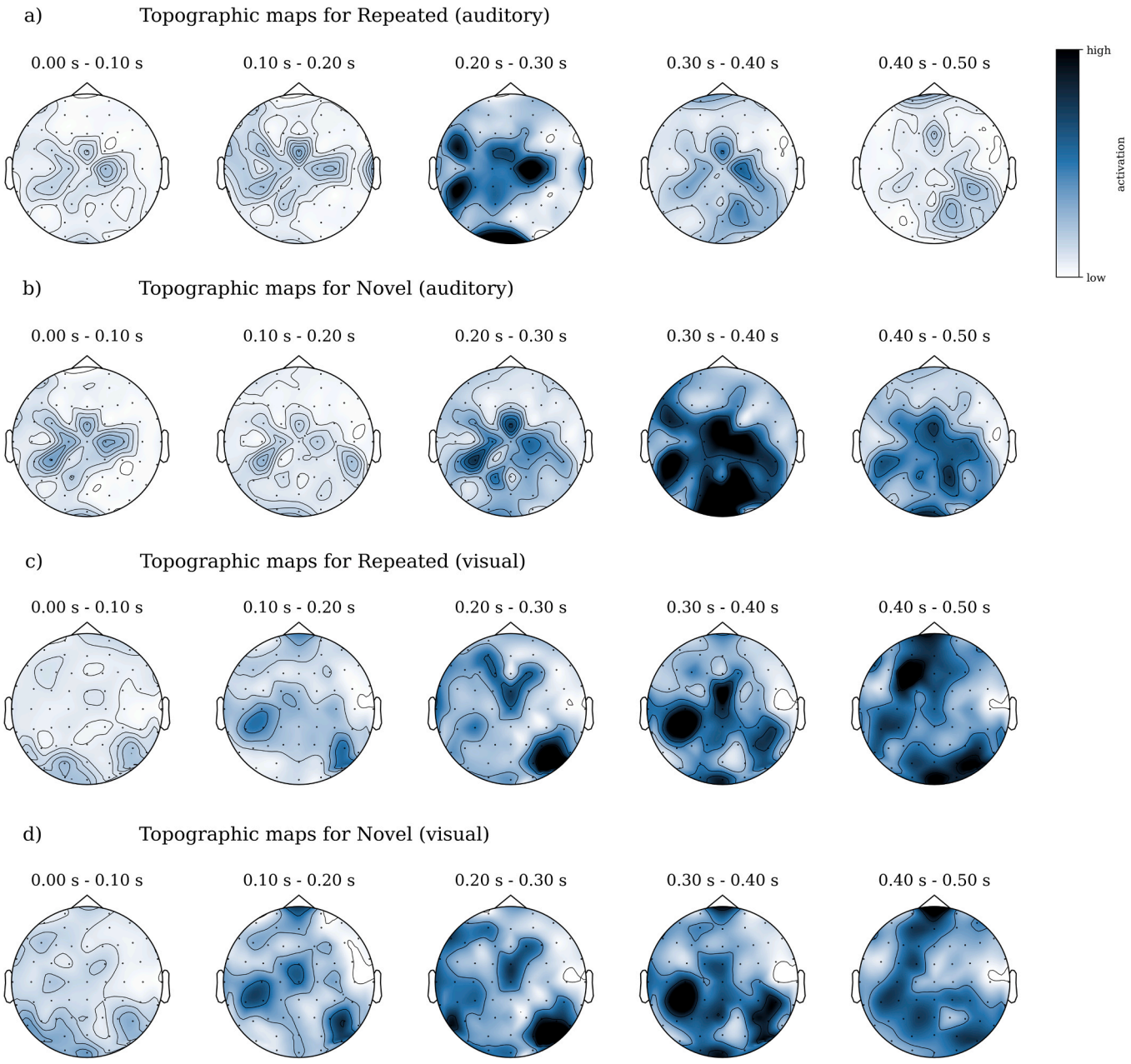

Fig. 10. Activation maps from figure 9 represented as topographic maps, across datasets and experimental conditions. For reasons of consistency, the topographic activation maps are displayed at the subset of commonly available latencies across the two datasets (i.e. 0–0.5 s post-stimulus onset). Every map corresponds to the sum of activations over intervals of 0.1 s.

did not have any periods of significant difference. As an illustration, the trial by trial activations extracted with activation maps and with the raw EEG data at the point of maximum regression (0.206 s post stimulus onset), are displayed in Fig. 11, panels c and d respectively.

4. Discussion

We presented a novel MVPA pipeline for decoding single trial EEG responses to external stimuli and used this pipeline to extract discriminant features at the single trial level. We showed, in two different datasets, that the developed pipeline significantly outperformed commonly used existing MVPA techniques, and that it could detect class-specific discriminant features that are readily interpretable. Our approach resulted in an accurate decoding performance, demonstrated in several ways: (a) generalization of classification to data that the network has not seen during training (test set), (b) generalization of the training pipeline to a different dataset (visual dataset), (c) significantly better classification performance for the original data vs. data with shuffled labels, (d) significantly higher classification performance for the network compared to exemplar baseline machine learning algorithms, and (e) comparable decoding performance to existing CNN-based algorithms for EEG data. Additionally, we used feature visualization techniques to characterise the electrodes and time-periods of EEG responses that mostly contribute to an accurate classification. Although several multivariate decoding techniques allow the extraction of discriminant features (Tzovara et al., 2012; Grootswagers et al., 2017), these are typically identified at an across trial level and are shared across experimental conditions. By contrast, our approach allows recovering class- and trial- specific discriminant features, which are informative of the distinct contributions of different experimental conditions to the

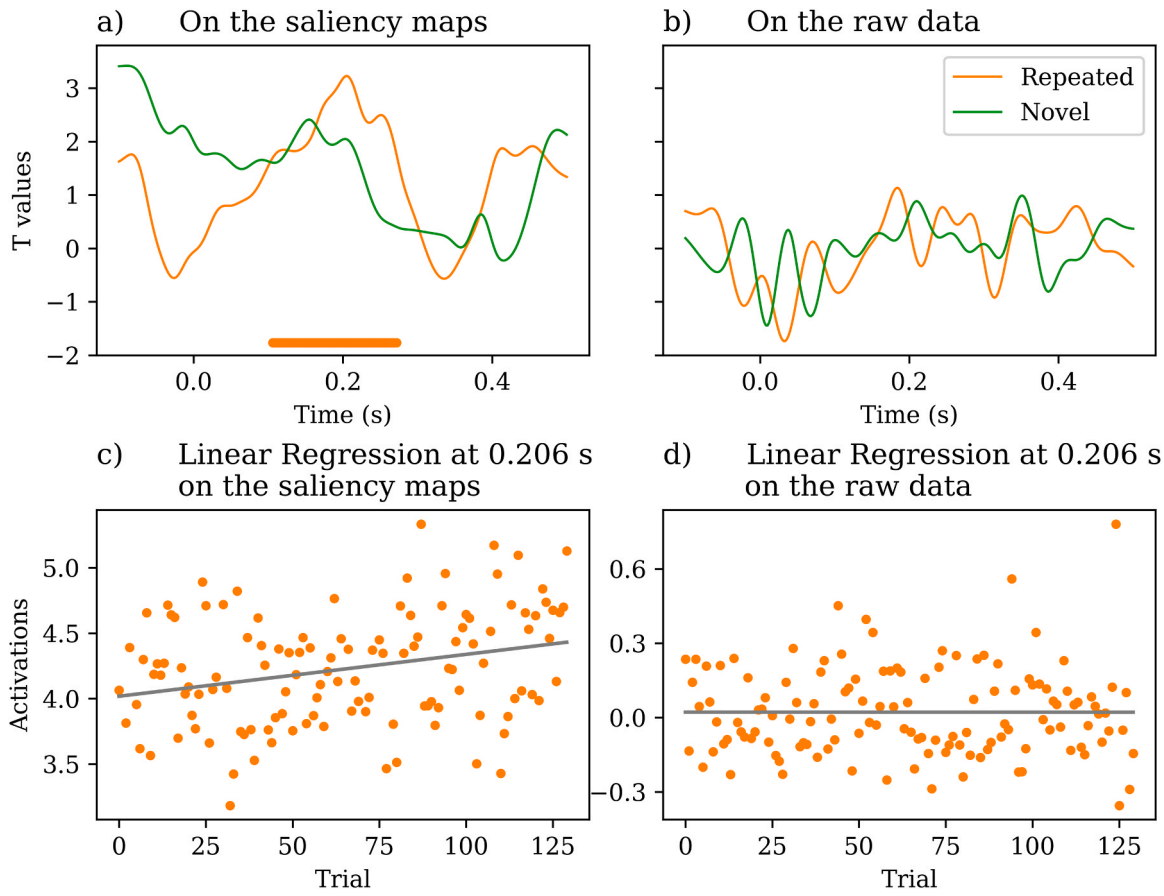

Fig. 11. Linear regression results quantifying trial by trial changes in the activation maps (panels a and c) and the raw data (panels b and d). The plotted lines correspond to the t-values testing whether the slope of a linear regression was significantly different from zero, for repeated (orange) and novel (green) stimuli. Horizontal lines show periods of significant difference. Panels c and d show trial by trial activations (panel c) or EEG responses (panel d) and regression fit, for the time point (0.206s) with the minimal p-value (0.002) for the *Repeated* condition.

final classification.

4.1. Convolutional neural networks for MVPA on EEG data

MVPA algorithms have been used to extract patterns of discriminant activity at the single trial level (Lemm et al., 2011; Haufe et al., 2014). The vast majority of these algorithms require a homogeneous dataset as they assume that the discriminant EEG responses appear at consistent latencies and electrode locations across trials (Grootswagers et al., 2017). However, often the discriminant information can be found at different temporal and spatial points over a group of participants or it may shift in time or space over the course of an experiment. The rather conservative approach of most multivariate decoding techniques can result in under-estimation of discriminant features, therefore limiting their interpretability. As CNNs convolve the entire input signal with multiple kernels per layer, they can extract patterns of neural activity which are time- and space- unlocked. Here, we chose a ResNet50 architecture, which is well known for its breakthrough performance in image classification in computer vision He et al. (2016). The depth of the network allows it to learn features and find structure in bigger patches of the input data than more shallow convolutional networks (Zeiler and Fergus, 2013). Even tough the network was originally developed for classifying images, here we adapted it for the specific case of EEG data.

One main concern of implementing deep learning algorithms for the field of EEG research is that of overfitting the data (Williams et al., 2020). Deep learning architectures typically comprise of hundreds to thousands of hyperparameters, which are prone to overfitting (Srivastava et al., 2014). Data augmentation techniques have been widely used in the field of computer vision (Shorten and Khoshgoftaar, 2019) to overcome this problem. This is a major concern for MVPA, as due to the nature of EEG recordings, it is practically impossible to collect the amounts of data that are often available in computer vision. Here, we overcame this limitation by using data augmentation techniques, which artificially augment the available EEG data and at the same time add variance to them, which makes the network less prone to overfitting to the available data samples. We could exclude that the trained networks were overfitted by showing that they generalize to new data (validation and test datasets). By contrast, the networks trained on randomly shuffled labels did not generalize to test and validation datasets.

The neural networks were trained for a maximum of 50 epochs (Fig. 4) and the training performance reached a plateau around epoch 15 with a mean score of 0.89 and 0.82 for the auditory and visual datasets respectively. The lower classification performance for the visual dataset was likely due to the nature of the visual Event Related Potentials that resulted in subtle differences between the two experimental conditions (van Peer et al., 2017). The *Repeated* and *Novel* visual stimuli were counter-balanced across participants, and therefore resulted in similar visual features at the average level, where the only difference would be a very subtle difference of the effect of repetition (see also Fig. 1). By contrast, in the auditory dataset (Cavanagh et al., 2018), repeated and novel sounds were always the same across participants and had very different acoustic characteristics (pure tones vs natural sounds). Therefore, for the auditory dataset, EEG responses were well distinct (Fig. 1), which resulted in a high classification performance.

When randomly shuffling the labels of the data to estimate a data-driven distribution of chance, the network accuracy could not

overcome 0.7 on average for the train data and remained at 0.5 for the test set (Fig. 7). Taken together, these results suggest that the network was able to learn class-specific features at an above-chance level without overfitting the data.

4.2. Comparison with existing techniques

Deep learning techniques have been introduced in the field of EEG research since a few years, but existing techniques predominantly focus on clinical or brain computer interface applications (Roy et al., 2019). Although deep learning techniques for basic EEG research exist (Kuanar et al., 2018; Bashivan et al., 2016; Wang et al., 2018), these are still at a validation stage, and are seldom used to answer basic research questions. Here, we aimed at examining how deep learning architectures perform in classification tasks that are commonly faced in basic neuroscience, i.e. decoding differences between experimental conditions, and evaluating the stability of decoding features over the course of an experiment. Indeed, the techniques that are readily available for classifying EEG data oftentimes give little to no emphasis on feature interpretability, but rather focus on classification performance (Williams et al., 2020). Although optimizing classification performance is certainly beneficial in basic research, clinical and BCI applications of MVPA algorithms can also profit from interpretable features. Importantly, our approach gave comparable results to other CNN-based architectures for classifying EEG data, based on a *Shallow* and *Deep* CNN (Schirrmeister et al., 2017).

Previous studies examining features of CNNs have extracted spectral EEG features (Schirrmeister et al., 2017), which do not contain temporal information, but are collapsed over time and trials. Such an approach is particularly suited for the field of brain computer interfaces, where emphasis is given on classification performance, but it is limited for MVPA applications, where emphasis is given on identifying spatial and temporal features that drive an accurate classification (Grootswagers et al., 2017). Other attempts to extract class-specific features based on gradient methods have either reported features at an average level (Farahat et al., 2019; Vahid et al., 2020), or for exemplar trials (Lawhern et al., 2018), without examining how these generalize over time, or how representative they are of the entire dataset. In Fig. 9 we also show the features on an average level, but additionally we examined how these change over the course of an experiment. Farahat et al. (2019) and Lawhern et al. (2018) both explore the features in the context of BCI and Vahid et al. (2020) and Lawhern et al. (2018) impose to the used network to start by temporal convolutions followed by spatial ones. Here, instead we consider the EEG data as a spatio-temporal continuum, as it is commonly done in the field of EEG research Maris and Oostenveld (2007).

4.3. Extraction of class-specific discriminant features

In computer vision, there is a dedicated research area focusing on visualising which features of the input data are learned by a neural network. Some commonly used techniques consist of visualising the kernels of the network (which provide only general features for the entire dataset that was used for training), or of gradient-based methods, which backpropagate the input signal trough the network (which can reveal class-specific features for each data-point). In neuroscience, there have recently been some studies focusing on feature interpretability (Ghosh et al., 2018; Lawhern et al., 2018; Schirrmeister et al., 2017; Zubarev et al., 2019). Here, we visualise discriminant features with a gradient-based method. The mean activation maps across participants (Figs. 9 and 10) showed similar spatio-temporal patterns of differential activity as the mean event-related potentials (Figs. 1 and S1).

The advantage of the presented pipeline for MVPA applications is that it allows to identify discriminant features at the single class level. Most existing univariate or multivariate analysis techniques can only identify condition differences, and are agnostic to which experimental condition is driving these differences. Other approaches, like common spatial patterns (CSP), allow the extraction of class specific patterns of EEG activity. However, the CSP components are calculated over a time window and therefore have a poor temporal resolution. Additionally, CSP have the limitation of poor generalization over new participants (Reuderink and Poel, 2008).

With our approach, in the auditory dataset, the strongest activation values were most prominent around 0.2 s post-stimulus onset for the *Repeated* condition and at later latencies, after 0.3 s post-stimulus onset for the *Novel* condition (Fig. 9 panels a and b). This difference in discriminant intervals for the two conditions is justified by the nature of the auditory stimuli, as *Repeated* sounds were pure tones with a sharp onset time, while *Novel* sounds were naturalistic sounds, which typically have a slower onset, and thus are expected to evoke an EEG response at later latencies (Cavanagh et al., 2018). This information cannot be revealed by existing MVPA techniques, which can only identify at which latency multiple conditions differ. Indeed, our analysis on the same data with an exemplar MVPA approach showed that classification was significantly higher than chance starting after 0.1 s post-stimulus onset, with a prominent peak after 0.3 s (Fig. 6). However, it is impossible to infer which of the two conditions drives this sustained differential activity. For the visual dataset, the peak in discriminant activity between *Repeated* and *Novel* stimuli appeared at latencies which were qualitatively similar between the two experimental conditions (Fig. 9 panels c and d). Indeed, in this dataset, the *Novel* and *Repeated* stimuli were all naturalistic images, and were counterbalanced across participants, therefore resulting in similar sensory responses (van Peer et al., 2014). In accordance to previous reports using this dataset, we found that the most prominent differences occurred after 0.1 s post-stimulus onset, and were sustained mostly up to 0.75 s, but also throughout the trial (Fig. 9 panels c and d). Importantly, when visualized in the form of topographic maps, the discriminant features that were identified through the CNNs match previous reports of this dataset, showing that topographic differences in response to novelty are mainly localized in frontal electrodes (van Peer et al., 2014). For the auditory dataset, the most prominent discriminant features at the topographic level were captured between 0.3 and 0.4 post-stimulus onset for the novel condition (Fig. 10). This latency and electrode locations are in accordance with previous reports of this dataset and could reflect a P3a component in response to novelty (Cavanagh et al., 2018).

To highlight the importance of studying discriminant features, we show that activation maps significantly change over the course of the experiment (subsection 2.7, Fig. 11). The positive t-values suggest that there was an increase in network activations over the course of the experiment, consistently observed between 0.106 and 0.272 s, suggesting that EEG responses at this latency activate neurons of the CNN more in later trials of the experiment compared to earlier ones. This could not have been caused by a global change in the strength of neural activity itself as our control analyses on raw EEG data did not find any significant trial by trial changes. Instead, our findings suggest that this increase might be caused by changes in the activation patterns in response to repeated stimuli, which become more distinct across conditions as the experiment unfolds. This interpretation is further supported by the fact that activations increase over the course of the experiment, as participants are increasingly exposed to the presented stimuli, and the two classes (*Repeated* vs. *Novel*) start acquiring a more distinct neural representation. Indeed, the observed latency for changes in activation maps is in accordance to the latency of a typical N100 response to auditory stimuli, which is known to habituate with repeated stimuli presentations (Rentzsch et al., 2008).

Studying changes of discriminant features can be relevant not only for basic EEG research, but also for BCI or clinical applications. In BCI applications it is highly relevant to investigate feature interpretability and how these features may manifest or change over long experimental sessions, as this may be relevant to participants' capacity to control an external device (Friedrich et al., 2013). Similarly, studying feature

interpretability in clinical applications is particularly important, in order to advance our understanding of which features may underlie algorithmic decisions, which in turn can contribute in gaining novel insights into neurological disorders.

4.4. Future directions and limitations

Currently, in the field of EEG research there is a lack of data-driven approaches that can provide information about trial-by-trial changes in EEG responses to external stimuli. Previous studies investigating, for example, learning of new sensory rules, have defined a priori a specific electrode locations and latencies of a response of interest and have examined how these change across trials (Lieder et al., 2013). Here, we refer to an alternative approach, that is 'data-driven' in the sense that features are identified automatically from the data, via means of maximizing discrimination between conditions of interest (i.e. supervised learning), as opposed to a hand-crafted feature selection that relies on a priori hypotheses (Haynes and Rees, 2006). Although the latter approach has been widely used in the literature of basic EEG research, it assumes that the response of interest stays at the same electrode location and latency across all trials, which in cases of changing processes, may not be true. Here, we propose a data-driven approach for identifying discriminant patterns of activity at the single-trial level and show that it is more sensitive than considering raw EEG activity (Fig. 11). Future studies can apply this approach in experiments involving learning in order to couple changes in discriminant EEG activity with participants' behavior and test learning theories.

One main limitation in the extraction of discriminant features is that they only reveal changes in network activations, but not the cause of these changes. Future experiments could evaluate changes of the network activation within spatial clusters (as identified in figure S3), possibly combining those with inverse solutions (He et al., 2006), in order to evaluate the contribution of specific brain regions in significant network activation changes over the course of an experiment (Lracitano et al., 2021).

Another future direction of research is that of choosing a network architecture. Here, we chose ResNet50 as an exemplar residual CNN, which has been widely tested in the field of computer vision (He et al., 2016) and biomedical data analysis (Guo and Yang, 2018). This choice is meant as a proof of principle, that demonstrates the feasibility of applying CNNs on MVPA applications for EEG data. Indeed, when changing the network architecture with other CNNs that have been developed for EEG research, classification performance remained at similar levels. Future studies can test different network implementations, to optimize the architecture and range of parameters for specific experimental setups. Additionally, here we chose to focus on a binary classification problem, for reasons of simplicity. Our present pipeline, as well as future attempts, could be easily expanded for multiple classes.

5. Conclusions

In summary, we used deep learning techniques to develop a novel MVPA pipeline for EEG data. We showed, in two different datasets, that our pipeline can accurately classify single-trial EEG responses, outperforming existing MVPA approaches, and performing at comparable levels with other deep learning approaches for EEG. Moreover, the neural networks can detect class specific information and discriminant features at the single trial level, a direction that can be used in the future to test theories of learning in a data driven way.

CRediT authorship contribution statement

Florence Aellen: Methodology, Software, Formal analysis, Data curation, Writing – original draft. **Pinar Göktepe-Kavis:** Data curation. **Stefanos Apostolopoulos:** Methodology, Supervision. **Athina Tzovara:** Conceptualization, Methodology, Writing – review & editing, Supervision.

Decleration of Competing Interest

Authors have no conflict of interest to disclose.

Data availability

The data used in this manuscript were taken from public repositories, and were originally referred to: Cavanagh, J. F., Napolitano, A., Wu, C., and Mueen, A. (2017). The patient repository for eeg data + computational tools (predict). Frontiers in Neuroinformatics. van Peer, J. M., Coutinho, E., Grandjean, D., and Scherer, K. R. (2017). Emotion-antecedent appraisal checks: Eeg and emg datasets for novelty and pleasantness. https://doi.org/10.5281/zenodo.197404.

Acknowledgements

This work is supported by the Interfaculty Research Cooperation 'Decoding Sleep: From Neurons to Health & Mind' of the University of Bern, Switzerland, the Swiss National Science Foundation (#320030_188737 to A.T.) and the Fondation Pierre Mercier pour la science (Switzerland).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jneumeth.2021.109367.

References

An, X., Kuang, D., Guo, X., Zhao, Y., He, L., 2014. A deep learning method for classification of eeg data based on motor imagery. Intell. Comput. Bioinformatics 203–210.

Bashivan, P., Rish, I., Yeasin, M., and Codella, N. (2016). Learning representations from eeg with deep recurrent-convolutional neural networks.

Burrello, A., Schindler, K., Benini, L., Rahimi, A., 2020. Hyperdimensional computing with local binary patterns: one-shot learning of seizure onset and identification of ictogenic brain regions using short-time ieeg recordings. IEEE Trans. Biomed. Eng. 67 (2), 601–613.

Castegnetti, G., Tzovara, A., Khemka, S., Melin, F., Barnes, G., Dolan, R., Bach, D., 2020. Representation of probabilistic outcomes during risky decision-making. Nat. Commun. 11.

Cavanagh, J.F., Kumar, P., Mueller, A.A., Richardson, S.P., Mueen, A., 2018. Diminished eeg habituation to novel events effectively classifies parkinson's patients. Clin. Neurophysiol. 129 (2), 409–418.

Cho, K.-O., Jang, H.-J., 2020. Comparison of different input modalities and network structures for deep learning-based seizure detection. Sci. Rep. 10.

Demarchi, G., Sanchez, G., Weisz, N., 2019. Automatic and feature-specific prediction-related neural activity in the human auditory system. Nat. Commun. 10.

Farahat, A., Reichert, C., Sweeney-Reed, C.M., Hinrichs, H., 2019. Convolutional neural networks for decoding of covert attention focus and saliency maps for eeg feature visualization. bioRxiv.

Fauw, J., Ledsam, J., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., Driessche, G., Lakshminarayanan, B., Meyer, C., Mackinder, F., Bouton, S., Ayoub, K., Chopra, R., King, D., Karthikesalingam, A., Ronneberger, O., 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat. Med. 24.

Fchollet (2016). Github deep learning models.Accessed: 2010–09-30.

Fiorillo, L., Puiatti, A., Papandrea, M., Ratti, P.-L., Favaro, P., Roth, C., Bargiotas, P., Bassetti, C.L., Faraci, F.D., 2019. Automated sleep scoring: a review of the latest approaches. Sleep. Med. Rev. 48, 101204.

Friedrich, E.V., Scherer, R., Neuper, C., 2013. Long-term evaluation of a 4-class imagery-based brain-computer interface. Clin. Neurophysiol. 124 (5), 916–927.

Ghosh, A., dal Maso, F., Roig, M., Mitsis, G.D., and Boudrias, M.-H. (2018). Deep semantic architecture with discriminative feature visualization for neuroimage analysis.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press,.

Gratton, G., Coles, M., Donchin, E., 1983. A new method for off-line removal of ocular artifact. Electroencephalogr. Clin. Neurophysiol. 55 (4), 468–484.

Grootswagers, T., Wardle, S.G., Carlson, T.A., 2017. Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. J. Cogn. Neurosci. 29 (4), 677–697. PMID: 27779910.

Guo, S., Yang, Z., 2018. Multi-channel-resnet: an integration framework towards skin lesion analysis. Inform. Med. Unlocked 12, 67–74.

Hajinoroozi, M., Mao, Z., Lin, Y.-P., and Huang, Y. (2017). Deep transfer learning for cross-subject and cross-experiment prediction of image rapid serial visual presentation events from eeg data.
Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., Bießmann, F., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage 87, 96–110.
Haynes, J.-D., Rees, G., 2006. Decoding mental states from brain activity in human. Nat. Rev. Neurosci. 7, 523–534.
He, B., Hori, J., Babiloni, F., 2006. Electroencephalography (EEG): Inverse Problems. American Cancer Society.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. 2016 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR).
Ieracitano, C., Mammone, N., Hussain, A., Morabito, F., 2021. A novel explainable machine learning approach for eeg-based brain-computer interface systems. Neural Comput. Appl. 1–14.
Jonas, S., Rossetti, A.O., Oddo, M., Jenni, S., Favaro, P., Zubler, F., 2019. Eeg-based outcome prediction after cardiac arrest with convolutional neural networks: performance and visualization of discriminative features. Hum. Brain Mapp. 40 (16), 4606–4617.
King, J.-R., Dehaene, S., 2014. Characterizing the dynamics of mental representations: the temporal generalization method. Trends Cogn. Sci. 18 (4), 203–210.
Kingma, D.P. and Ba, J. (2014). Adam: A method for stochastic optimization.
Koles, Z., Lazar, M., Zhou, S., 1990. Spatial patterns underlying population differences in the background eeg. Brain Topogr. 2 (4), 275–284.
Kotikalapudi, R. and contributors (2017). keras-vis.⟨https://github.com/raghakot/keras-vis⟩.
Kuanar, S., Athitsos, V., Pradhan, N., Mishra, A., Rao, K.R., 2018. Cognitive analysis of working memory load from eeg, by a deep recurrent neural network. 2018 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP) 2576–2580.
Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J., 2018. Eegnet: a compact convolutional neural network for eeg-based brain-computer interfaces. J. Neural Eng. 15 (5), 056013.
Lemm, S., Blankertz, B., Dickhaus, T., Müller, K.-R., 2011. Introduction to machine learning for brain imaging. NeuroImage 56 (2), 387–399 (Multivariate Decoding and Brain Reading).
Lieder, F., Daunizeau, J., Garrido, M.I., Friston, K.J., Stephan, K.E., 2013. Modelling trial-by-trial changes in the mismatch negativity. PLoS Comput. Biol. 9 (2), 1–16.
Lotte, F. (2014). A tutorial on eeg signal processing techniques for mental state recognition in brain-computer interfaces.
Macmillan, N., Creelman, D., 2004. Detection Theory: A Useras Guide, xix. Psychology Press,.
Maris, E., Oostenveld, R., 2007. Nonparametric statistical testing of eeg- and meg-data. J. Neurosci. Methods 164 (1), 177–190.
Mehrer, J., Spoerer, C.J., Kriegeskorte, N., Kietzmann, T.C., 2020. Individual differences among deep neural network models. Nat. Commun. 11 (1), 5725.
Nurse, E., Mashford, B.S., Yepes, A.J., Kiral-Kornek, I., Harrer, S., and Freestone, D.R. (2016). Decoding eeg and lfp signals using deep learning: Heading truenorth.
Philiastides, M.G., Biele, G., Vavatzanidis, N., Kazzer, P., Heekeren, H.R., 2010. Temporal dynamics of prediction error processing during reward-based decision making. NeuroImage 53 (1), 221–232.
Rentzsch, J., Jockers-Scherübl, M.C., Boutros, N.N., Gallinat, J., 2008. Test-retest reliability of p50, n100 and p200 auditory sensory gating in healthy subjects. Int. J. Psychophysiol. 67 (2), 81–90.
Reuderink, B. and Poel, M. (2008). Robustness of the common spatial patterns algorithm in the bci-pipeline.IEEE Transactions on Circuits and Systems I-regular Papers - IEEE TRANS CIRCUIT SYST-I.
Roy,, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T.H., Faubert, J., 2019. Deep learning-based electroencephalography analysis: a systematic review. J. Neural Eng. 16.
Schirrmeister, R.T., Springenberg, J.T., Fiederer, L.D.J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., Ball, T., 2017. Deep learning with convolutional neural networks for eeg decoding and visualization. Human Brain Mapp. 38 (11), 5391–5420.
Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. J. Big Data 6, 1–48.
Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.
Suzuki, K., 2017. Overview of deep learning in medical imaging. Radiol. Phys. Technol. 10.
Tan, C., Sun, F., and Zhang, W. (2018). Deep transfer learning for eeg-based brain computer interface.
Tang, Z., Li, C., Sun, S., 2016. Single-trial eeg classification of motor imagery using deep convolutional neural networks. Opt. - Int. J. Light Electron Opt. 130.
Tomioka, R., Aihara, K., Müller, K.-R., 2006. Logistic regression for single trial eeg classification. Proceedings of the 19th International Conference on Neural Information Processing Systems, NIPS'06. MIT Press, Cambridge, MA, USA, pp. 1377–1384.
Tzovara, A., Murray, M.M., Plomp, G., Herzog, M.H., Michel, C.M., De Lucia, M., 2012. Decoding stimulus-related information from single-trial eeg responses based on voltage topographies. Pattern Recognit. 45 (6), 2109–2122 (Brain Decoding).
Vahid, A., Mückschel, M., Stober, S., Stock, A.K., Beste, C., 2020. Applying deep learning to single-trial eeg data provides evidence for complementary theories on action control. Commun. Biol.
van Peer, J.M., Coutinho, E., Grandjean, D., and Scherer, K.R. (2017). Emotion-antecedent appraisal checks: Eeg and emg datasets for novelty and pleasantness.10.5281/zenodo.197404.
van Peer, J.M., Grandjean, D., Scherer, K.R., 2014. Sequential unfolding of appraisals: Eeg evidence for the interaction of novelty and pleasantness. Emotion 14 (1), 51–63.
Wang, F., Zhong, S.-H., Peng, J., Jiang, J., Liu, Y., 2018. Data augmentation for eeg-based emotion recognition with deep convolutional neural networks. In: Schoeffmann, K., Chalidabhongse, T.H., Ngo, C.W., Aramvith, S., O'Connor, N.E., Ho, Y.-S., Gabbouj, M., Elgammal, A. (Eds.), MultiMedia Modeling. Springer International Publishing, Cham, pp. 82–93.
Williams, J.M., Samal, A., Rao, P.K., Johnson, M.R., 2020. Paired trial classification: a novel deep learning technique for mvpa. Front. Neurosci. 14, 417.
Zeiler, M.D. and Fergus, R. (2013). Visualizing and understanding convolutional networks.
Zhang, X., Yao, L., Wang, X., Monaghan, J., Mcalpine, D., and Zhang, Y. (2019). A survey on deep learning based brain computer interface: Recent advances and new frontiers.
Zubarev, I., Zetter, R., Halme, H.-L., Parkkonen, L., 2019. Adaptive neural network classifier for decoding meg signals. NeuroImage 197, 425–434.