On the use of Pairwise Distance Learning for Brain Signal Classification with Limited Observations

David Calhas

INESC-ID and IST, Universidade de Lisboa, Lisbon, Portugal

Enrique Romero

Universitat Politcnica de Catalunya, Barcelona, Spain

Rui Henriques

INESC-ID and IST, Universidade de Lisboa, Lisbon, Portugal

Abstract

The increasing access to brain signal data using electroencephalography creates new opportunities to study electrophysiological brain activity and perform ambulatory diagnoses of neurological disorders. This work proposes a pairwise distance learning approach for Schizophrenia classification relying on the spectral properties of the signal. Given the limited number of observations (i.e. the case and/or control individuals) in clinical trials, we propose a Siamese neural network architecture to learn a discriminative feature space from pairwise combinations of observations per channel. In this way, the multivariate order of the signal is used as a form of data augmentation, further supporting the network generalization ability. Convolutional layers with parameters learned under a cosine contrastive loss are proposed to adequately explore spectral images derived from the brain signal. With the proposed method achieving a best of 0.95 ± 0.05, 0.98 ± 0.02 and 0.92 ± 0.07 in terms of Accuracy, Sensitivity and Specificity respectively. Results on a case-control population show that the features extracted using the proposed neural network are superior than baselines to diagnose Schizophrenia, suggesting the existence of non-trivial electrophysiological brain patterns able to capture discriminative neuroplasticity profiles among individuals.

Keywords:
Pairwise Learning, Schizophrenia, Classification, Electroencephalography

*Corresponding author at: INESC-ID, Rua Alves Redol 9, 1000-029, Lisboa, Portugal
Email addresses: `david.calhas@tecnico.ulisboa.pt` (David Calhas), `eromero@cs.upc.edu` (Enrique Romero), `rmch@tecnico.ulisboa.pt` (Rui Henriques)

1. Introduction

The recording of increasingly affordable electroencephalography (EEG) and precise data is creating unprecedented opportunities to understand brain activity, aid personalized prognostics, and promote health through wearable biofeedback systems [1]. Electroencephalography is non-invasive, safe, inexpensive, and shows rich temporal content; in contrast with other brain imaging modalities, such as magnetic resonances, entailing higher costs and restrictions on the periodicity of recordings [2]. EEG monitoring is widely used to assess psychiatric disorders, and has shown to be a valuable source to study Schizophrenia, a disorder affecting about 1% of the world population, largely susceptible to misdiagnoses [3].

In [4], a neurofeedback training was performed on a female patient, who suffered from schizophrenia for more than 7 years and had had several schizophrenic episodes. This type of training enables the individuals to regulate their brain activity using a real-time feedback loop. After the neurofeedback training, increased amplitude in alpha waves and decreased amplitude in beta waves were observed. The patient found out the most effective mental strategies and learnt how to regulate her brain activity (mental strategies were induced with the help of a psychotherapist, being the most effective mental strategies: natural scenery including waterfall, moon, lake, mountain, lotus and sea). In [5], a status overview on EEG abnormalities is given, in order to diagnose patients with schizophrenia. To this end, they examined the status of development of spectral EEG deviations. In the gathered studies, the meta analysis was limited to those works comparing spectral power between one group of schizophrenia patients and one group of healthy control subjects. The presence of two groups (or populations), one with the pathology and a healthy control group, is essential to identify discriminative features from the gathered signals. The hypothesized differences of schizophrenia individuals were increased delta, increased theta, decreased alpha, and increased beta power. A number of subsequent studies suggested that an increase of activity in the lower spectrum (slow waves) is significantly higher in schizophrenia populations. It is also noted that slow wave abnormality (mainly delta increase) is mainly localized in frontal lobe regions. One of the conclusions, is that the delta excess (and to a lesser extent the theta excess) is a strong biological marker of schizophrenia [5]. Event-related EEG stems suggest that neural oscillations and their synchronization represent important mechanisms for interneuronal communication and binding of information that is processed in distributed brain regions [6]. Several studies with EEGs have in mind the gain of new insights into the pathophysiological processes underlying cognitive deficits in neuropsychiatric disorders, supporting the role of EEG data analysis to study schizophrenia. Frequency analysis is hugely encouraged to do when performing a schizophrenia study with EEG. This is verified by comparing our proposed model and other frequency related methods with Event Related Potentials based models [7, 8, 9], that purely process the signal in a time domain not taking into account the frequency domain.

Despite the inherent advantages of monitoring electrophysiological brain activity, its use for diagnosing neuronal diseases is still capped by the limited size of case-control populations [10], as well as by the intrinsic difficulties of mining brain signals. Brain signal data is high-dimensional, multivariate, susceptible to noise/artefacts, rich in temporal-spatial-spectral content, and highly-variable between individuals [11]. In this work one dataset is used [12], further details can be found in Section 3.1.

This work proposes a dedicated class of neural networks to extract discriminative features of Schizophrenia from electrophysiological brain data. The proposed approach combines principles from pairwise distance learning and spectral imaging in order to address the aforementioned challenges, enabling superior diagnostics. Accordingly, the proposed approach offers six major contributions:

1. Ability to learn from small datasets by taking advantage of Siamese Network layering, inherently prepared to work in augmented data spaces mapped from a limited number of observations (the dataset employed only has 84 instances [12]). The features produced by these networks have proven to be useful to perform classification as they rely on either the homologous or discriminative properties of observation-pairs in a pairwise distance domain [13];
2. Ability to deal with the rich and complex spectral and temporal content of EEG data by processing the signal into spectral images with a fine frequency and temporal resolution per electrode [14, 15], and by subsequently reshaping the Siamese network architecture with adequate convolutional operations;
3. Robustness to noise and wave-instability by assessing distances on the spectral content under a cosine-loss. Gathered evidence shows less susceptibility to artefacts and the inherent variability of electrophysiological potentials associated with continuously changing overlapping electrical fields produced by localized neurons [11];
4. Ability to deal with the multivariate nature of the signal (rich spatial content) by capturing interdependencies between channels as their content is simultaneously used to shape the weights of shared connections in the network;
5. Ability to handle the extremely-high dimensional nature of the gathered spectral content from brain signals (high-resolution spectral image per electrode) under L1 regularization [16, 17];
6. Applicability of the proposed EEG-based diagnostics to alternative populations or diseases, evidenced by the: i) placed Bayesian optimization step [18] for hyperparameter tuning and fixing feature numerosity; ii) fully-automated nature of the approach once signals are recorded; and iii) generalization ability of the learning process on validation data.

In contrast with the traditional neural information processing systems, this manuscript explores whether we can go deep on highly-dimensional spatiotemporal data in the presence of a very limited number of data observations. This stance is much needed in healthcare given the limited size of trials (cohort studies), often driven by disease rarity, capped size of control population, trial eligibility requirements, or the facultative nature of EEG assessments. Results confirm this possibility: +20pp in the accuracy and sensitivity of Schizophrenia diagnostics.

The features extracted from the proposed spectral and pairwise distance space further suggest the presence of discriminative elecrophysiological patterns linked to neuroplasticity aspects of the individuals. This observation is in accordance with findings from previous studies that established statistically significant relationships between variations in the frequency band spectrum and neuroplasticity conditions [19, 20].

The manuscript is organized as follows. After formalizing the problem, Section 2 surveys existing contributions on the diagnosis of individuals from brain signal data. Section 3 describes the proposed solution. Section 4 shows extended evidence of its relevance for diagnosing Schizophrenia. Finally, concluding remarks are drawn in Section 5.

1.1. Problem formulation

Problem. A EEG recording or brain signal observation is a multivariate time series $X{=}\{x_t^j \mid j \in \{1..M\}, t \in \{1..T\}\}$, where x_t^j is a measure of the electrophysiological activity in scalp channel j and instant t, T is the number of time points, and M is the multivariate order (number of channels). Given brain signal dataset, $\{(X_i, c_i) \mid i = 1..N\}$, where N is the number of EEG recordings and each recording X_i is annotated with a label $c_i \in \Sigma$, our task is to identify a discriminative feature space to classify (unlabeled) observations. Specifically, we are interested in classifying Schizophrenia given case-control populations.

Essential background. The electrophysiological signal produced by a specific channel in the cerebral cortex is a univariate time series that can be decomposed into a frequency time series using a discrete Fourier transform. The analysis of the frequency domain of a signal, generally referred as spectral analysis, determines the predominant waves monitored at a certain location. A short-time discrete Fourier transform can be alternatively applied along a sliding window of the raw signal to capture potentially relevant changes on the spectral activity of the brain throughout the EEG recording. The spectral content produced by this time-varying form of spectral analysis is here informally referred as a *spectral image* since it measures brain activity along two contiguous axes: time and frequency.

2. Related Work

Recent works on deep learning provide principles to attemptively learn from small datasets [21, 22], a critical requirement to guarantee their applicability for most cohort studies. The use of surrogate data analysis for regresssion tasks [21], or data augmentation procedures for image recognition [23] are paradigmatic cases. Despite their relevance, they either tackle different tasks or assume a substantial higher amount of data observations than the ones commonly available in cohort studies; leave aside the need to handle the high dimensionality, spectral variability, and rich spatiotemporal content of EEG data.

2.1. EEG Classification

EEGNet [8], EEGNet-SSVEP [8], DeepConvNet [9] and ShallowConvNet [9] are considered state-of-the-art EEG classification built models that make use of convolutional operations directly on the raw EEG data. These convolutions are placed along time and channels. Approaches like these rely on the properties of its models to extract discriminative features from EEG signals. These models have been primarily validated in the context of stimuli-induced recording sessions. One can see directly that these networks learn event related potentials from the EEG signal, which makes the EEG recording session dependable of a task environment for evoking potentials. In contrast, we aim at extracting neuroplasticity-related features from resting state EEG data, for which effective deep learning methods are still in demand. Section 4 confirms the limited relevance of existing methods to learn from resting state EEG data.

2.2. EEG on Schizophrenia

Dvey-Aharon et al. [24] claim mostly changes in functional connectivity are seen in patients with Schizophrenia, as well as differences in theta-frequency activity. A classification approach was applied on 1-minute signals recorded by a single electrode. The developed system consists of four stages: performing several preprocessing tasks and breaking the raw signals into relevant intervals; transformation of the EEG signal into a time/frequency representation via the Stockwell transformation; feature extraction from the time/frequency representation; and discrimination of specific time frames following a given set of stimuli between the time/frequency matrix representations of the healthy subjects and the schizophrenia patients. With this, for each subject a window (of 1.2 seconds, before and after each stimulus) was taken from the EEG signal, and filtered for a set of frequencies (composed of the average of the subjects signals according to a parameter). For each interval Stockwell features were extracted, producing a feature vector that is classified according to the K-Nearest Neighbor classifier using the euclidean distance. Despite promising results, the approach requires the performance of cognitive tasks by the individuals under assessment throughout the recording. More recently, the authors introduced another way of looking at the EEG signal using connectivity maps derived from the brain activity [25]. In order to build these maps, a similarity function needs to be chosen, so one can check which nodes are more similar to which ones. Results showed that the degradation of connectivity is being accelerated within schizophrenia individuals. And that information relay changes in an abnormal manner primarily in the prefrontal area. This gives a good insight on how connectivity maps can be applied to discriminate schizophrenia. And most important, that one should take into account that a change in a certain region can influence other regions in the brain.

Sabeti et al. [26] introduced another approach to classify Schizophrenia based on entropy and complexity measures of the EEG signal. The features extracted from the signal were: Shannon entropy, spectral entropy, approximate entropy, Lempel-Ziv complexity and Higuchi fractal dimension. Genetic programming was used for feature selection. With these features, Adaptative boost (Adaboost) and Linear Discriminant Analysis (LDA) classifiers were validated, showing performance improvements against peer approaches. The recordings were done with eyes open, a setting easily biased by environmental effects.

Notable examples of connectionist and spectral approaches were introduced to discriminate and characterize Schizophrenia. Nevertheless, there is still a research gap on how to simultaneously explore the rich spectral, temporal and spatial nature of brain signals to perform classification. In spite of the indisputable role of neural network learning for the analysis of complex spatiotemporal signal data, its role for EEG-based diagnostics of psychiatric disorders remains largely unexplored due to the absence of large cohorts and the inherent stochastic complexities associated with electrophysiological data.

2.3. Deep Learning on EEG

Applying deep learning techniques on medical data has been challenging and despite the advances on data gathering, there continues to be a need for public data to be available [27]. Fortunately, recently published studies were able to tackle this problem given the lack of resources [28, 29].

[28] performs classification task with three classes: Alzheimers Disease, Mild Cognitive Impairment and Healthy Control. The classification is done by analyzing EEG recordings. The EEG recording session was setup with 19 channels and 189 recordings were gathered composing the three classes mentioned (63 individuals for each one). As stated by [28], standard machine learning methods are not able to deal with high dimensional data as it is the case for EEG data (taking into account channels and frequency bands). In contrast, Deep Learning techniques have shown the ability to perform feature selection (by extracting the more relevant ones) and thus tackle this problem. Convolutional operations were employed to extract features from the Power Spectrum Density, using ReLU as the activation function. One main difference of our work and [28] is that we take into account each frequency time series, in contrast they take the overall magnitude of each frequency. Analyzing each frequency time series is useful as it has been correlated with neuroplasticity properties of the brain.

[29] performs schizophrenia classification based on EEG recordings. The interesting particularity about this work is that it uses deep learning techniques, mainly convolutional operations. EEG recordings were gathered from 14 healthy controls and 14 schizophrenic individuals. A total of 19 electrodes were used and the signal was sampled at 250 Hz. The model was trained and then validated on the training data by a k-fold cross validation. [29] also employs convolutional operations on the time domain, which is not encouraged when following a resting state protocol. Fortunately the dataset used also had segments with

a naming activity (task oriented) and this may be the reason why the results were competitive (81.26% accuracy).

[28, 29] have a downside, which is the fact that the architecture chosen was manually tuned and can bring discussion on the nature of the hyperparameters chosen. In contrast, our work does not present any values for the hyperparameters of the network, we leave the tuning to the Bayesian Optimization algorithm. As such, a network with set values is not proposed, but the procedure to get to it is. The goal is to easily bridge the research done to a real world setting, where two populations of individuals are gathered and the hyperparameters are tuned by the Bayesian Optimization algorithm. After the procedure, the model is then able to diagnose patients quickly.

2.4. Siamese Neural Network

Siamese Neural Networks (SNN), first introduced by Bromley et al. [30] to distinguish signature forgeries from real ones, are deep learning architectures with two sub-networks that consist on the same instance, hence being called ”siamese networks”. This architecture receives as input a pair of samples. Subsequently, the outputs of the pairs used as input to these ”siamese networks” are joined in a distance function. The proposed distance function between the output of the SNNs is the cosine similarity (for signatures from the same person the output should be 1, and -1 for forged ones). This model had outstanding results at the time, detecting 80.0% of the forged signatures and 95.5% of the genuine signatures. More recently, Kock et al. [13] successfully used a SNN Architecture for One Shot Learning (meaning the model only sees each class once in an epoch). This approach reached 92.8% accuracy in the test set. These results were achieved through a Siamese Convolutional Architecture. Once this kind of network is trained, its learned representations via a supervised metric-based approach with SNNs are useful to perform tasks like classification, relying on the discriminative properties of these features.

Medicine diagnosis is based on analyzing symptoms and comparing them to the history (of patients with the same symptoms) in order to assign a class to a patient. One can say medicine has its roots on statistics [31], as it purely compares observations with history archives when making diagnosis. Our decision for the architecture employed was based on the nature of this field. In the case of a CNN for classification, it would learn the features overall that are able to discriminate between healthy controls and schizophrenic individuals, but there is no comparison necessarily being done. In contrast, an SNN learns by comparison and it is a motivation for its application in a medicine field task (schizophrenia diagnosis), due to the learning mechanism being quite similar to the way medicine diagnoses are operated by humans.

3. Our Approach

The proposed architecture is inspired by the architecture formerly introduced by Kock et al. [13]. An advantage of this type of architecture is the ability to augment the original dataset from an instance-based data space to a pair-based one. Our approach has two main steps: 1) feature extraction; and 2) classification. In step 1, the internal representations obtained from the SNN architecture model are extracted after training. In step 2, a classification task is performed using these extracted features. Previous to both steps, we perform hyperparameter optimization for every model using Bayesian Optimization (BO) [18].

3.1. Dataset Description

Approaches based on induced stimuli or task performance, followed by the analysis of event related potentials, are not considered in this work. Instead, a resting state setup is considered to monitor the underlying brain patterning at the brain cortex, independently of the surrounding environment/undertaken task. Subsequently, this avoids any additional interference on the EEG signal recorded. [32] findings support the use of this setup, claiming that differences on the spectral activity – such as higher delta and a lower alpha synchronization in psychotic disorders – can be optimally detected in resting state protocols with both open and closed eyes.

Table 1 shows the content of EEG datasets containing healthy control individuals and schizophrenic individuals. Dvey-Aharon et al. [25, 24] and Sabeti et al. [26] works were introduced and discussed in Section 2. Unfortunately, the considered datasets have a strictly low number of observations, and are not made publicly available. Nonetheless, Gorbachevskaya and Borisov [12] performed a broader resting state recording on a total of 84 individuals, of which 45 were schizophrenic and 39 were regarded as healthy controls. This dataset is publicly available and thus it was the one employed in this study. This population consists of adolescents who had been screened by a psychiatrist and got either a positive or negative diagnostic for the schizophrenia neuropathology. EEG recordings were sampled at 128 Hz with 1 minute duration. Individuals were set in a resting state with eyes closed. In accordance with the 10-20 system of electrode placement, the topographical positions of the placed EEG channels are: F7, F3, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1, O2.

Dataset reference	Healthy Controls	Schizohprenic Individuals	Access
Dvey-Aharon et al. [25, 24]	20	20	Private
Sabeti et al. [26]	25	25	Private
Gorbachevskaya and Borisov [12]	**39**	**45**	**Public**

Table 1: Schizophrenia EEG datasets.

3.2. Siamese Neural Network Architecture

The SNN architecture contains two sub networks that correspond to the same instance (twin networks). Both of these twin networks are referred to as the Base Network (BN). The input and output of the BN are an example and a feature vector, respectively. The output feature vector corresponds to the features extracted in the aforementioned step 1.

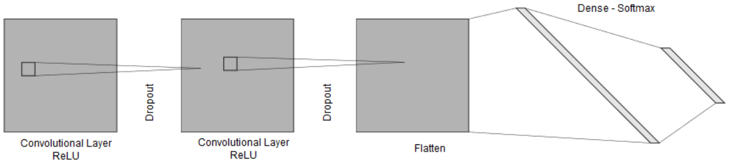


Figure 1: Base network from the SNN.

In our case, the BN receives as input a Discrete Short-Time Fourier Transform (DSTFT) representation of the EEG signal, that is extracted from the 1 minute recording of a channel of an individual. The DSTFT is taken with 2 seconds length windows in order to capture frequencies as low as 0.5Hz, corresponding to the delta wave frequencies (Howells et al. [32] points out that frequencies lower than 2Hz are relevant to differentiate Schizophrenia). This image is processed through two convolutional layers, followed by a fully connected layer. The activation function used in the convolutional layers is the rectified linear function [33], while the fully connected layer uses the softmax activation function, normalizing the domain of the feature representations, $\mathbf{f} \in R^q, i \in [1, q] : \mathbf{f}_i \in [0, 1]$.

Once the BN network (Fig. 1) is built, a replication of it is made, producing its twin and sharing their weights. The SNN layout is achieved joining these twins and computing a distance metric between their outputs, as shown in Fig. 2. In our case, the inputs to the SNN are pairs of DSTFT representations and the outputs are the computed distance between the representations obtained by the BN.

The SNN tries to solve what is known as a neighbor separation problem, consisting on the separation of instances in a dataset that contains different classes. In our case we have two classes: schizophrenic and healthy control individuals. In this neighbor separation problem, pairs of individuals of the same class (schizophrenic with schizophrenic or healthy with healthy) are called **neighbors** and pairs of individuals of different classes (schizophrenic with healthy) are called **non-neighbors**. The network learns a transformation with the objective of assigning small distance to neighbors and large distance to non-neighbors.

With the previously described architecture, the neighbor separation problem can be posed as a minimization problem of a certain loss function that depends on such distance. In [34], the Contrastive Loss function is introduced to that end, defined as:

$$L(W, Y, X_1, X_2) = Y {D_W}^2 + (1 - Y)\ max(0, m - D_W)^2 \tag{1}$$

where (X_1, X_2) is the input pair, $Y = 1$ if X_1 and X_2 are neighbors and 0 otherwise, D_W the distance between the predicted values of X_1 and X_2, and m is the margin value of separation. Minimization of the Contrastive Loss function leads to a scenario where neighbors are pulled together and non-neighbors are pushed apart, according to a certain distance metric. The margin value is sensitive. High values of m increase the separation between non-neighbors (pairs

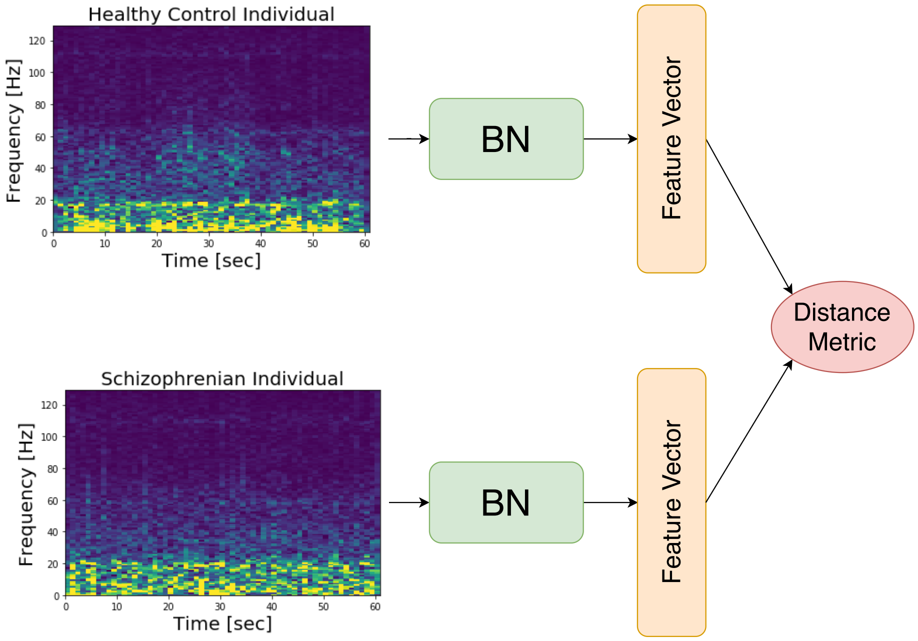


Figure 2: SNN architecture.

of different class), impacting positively the accuracy although making the training slower. In contrast, low values of m may cause the model not to learn the desired behavior.

3.2.1. Loss and Regularization

The suggested contrastive loss function to measure the correlation between two feature vectors is the cosine loss. This metric generally shows reasonable performance improvements, suggesting that the coherence of spectral variations between spectral images (cosine loss) is more relevant than the actual absolute differences between images (euclidean loss), an observation corroborated in other recent studies [22]. This observation also sheds light on how the schizophrenia pathology is expressed in the EEG.

Besides the type of layers and the distance metric, the following techniques are integrated in the model: $L1$ regularization and Dropout layers. The $L1$ regularization is useful because it helps remove features that are not useful for the task. Dropout layers are introduced to improve generalization. Regularization is applied at the kernel of all layers. The Dropout probability used is 0.5, as suggested by [35], and is applied after each convolutional layer. Adam [36] is used to optimize the network during the training session.

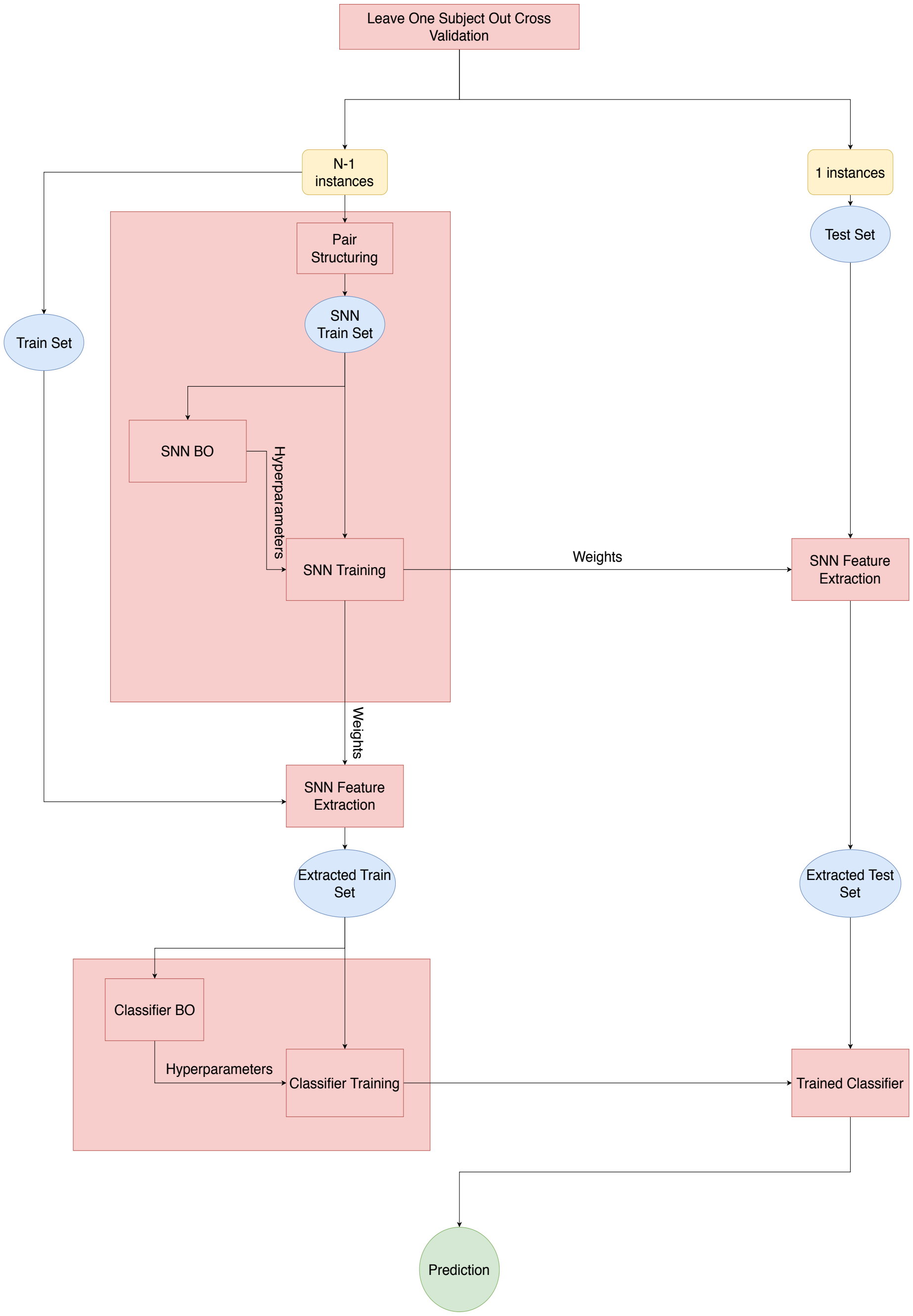


Figure 3: Schematic representation of the proposed validation procedure: SNN feature extraction (Section 3.2.2) and classification (Section 3.3).

3.2.2. Hyperparameter Tuning

The number of layers, as well as their type, are fixed. The rest of hyperparameters (regularization factor, margin value, learning rate, kernel size and output dimension of the BN) are susceptible to optimization. As previously mentioned, we apply BO to that end. BO is set to run with a maximum of 50 acquisitions and starts with 5 iterations to perform an initial exploration. In each iteration and acquisition, a K-fold Cross Validation with $K = 5$ is done with the training set of a Leave-One-Subject-Out Cross Validation (LOOCV) partition. The combination of hyperparameters that has the best average validation accuracy across the 5-folds is chosen to perform the feature extraction. Each of the hyperparameters are assigned the following value domains to explore: regularization factor $\in [10^{-3}, 10^{-1}]$, margin value $\in [1.0, 2.0]$, learning rate $\in [10^{-6}, 10^{-3}]$, kernel size $t \times f$ with $t = f = \{3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ (the same kernel size is used for both convolutional layers) and final output dimension $\in \{2, 4, 6, 8, 10, 12, 14\}$. The BO surrogate model is a standard Gaussian Process. Expected Improvement is used as an acquisition function and the Limited-Memory BroydenFletcherGoldfarbShanno algorithm as the acquisition optimizer.

The DSTFT magnitudes are normalized, under the hypothesis that there exists a threshold from which there is no additional information to identify the schizophrenia pathology. With this, the values are normalized by an upper value, U. Values of f smaller than U are divided by U and magnitudes bigger than U are set to 1.0. This allows every magnitude of the frequencies to be within the interval $[0, 1]$ after the normalization is performed. We take advantage of the BO exploration to obtain U, by introducing it in the same optimization process made for the SNN hyperparameters. The domain assigned to be explored for U is $[100.0, 500.0]$.

3.2.3. Pairwise Dataset Structure

To guarantee that the target network is able to learn valid transformation for all channels, the pairs are set such that only same channels are paired (Figure 4). Pairs of different channels are not considered, since different channels are seen as correlated spaces with different properties. Fortunately, the SNN is capable of learning different spaces/classes, as shown in [13], where the proposed system is able to learn a similar setup. The pairwise schema brings a new optimization space to the classifier and consequently more observations versus a traditional classifier, which is one of the strongest motivations to use the SNN architecture. The main difference is that from bringing the problem from a instance class problem to a relation classification one, it is transformed to a relation labeling one that ables us to have more pairs (much more than original instances) to learn from. The feature variance is the same, but the optimization space has more information available. Although no actual data augmentation scheme, such as image transformations (scaling, rotations) and noise addition, is applied, the number of instances increases and consequently the pairwise structured dataset is seen as a data augmentation from the standard dataset.

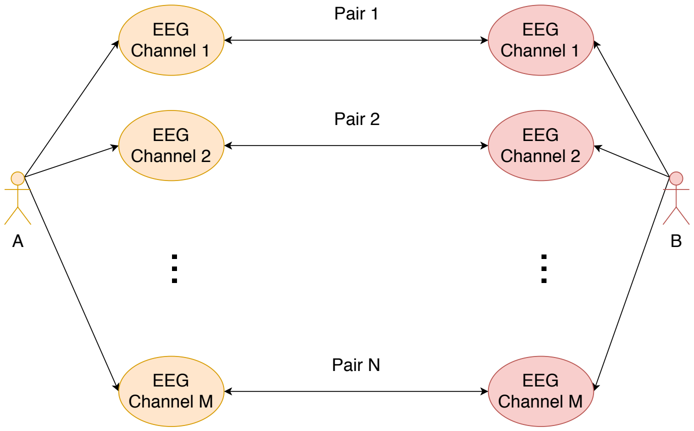


Figure 4: Pair Structure between two individuals and the corresponding EEG channels.

From our original EEG dataset, $X_1, ..., X_N$ spectral images are derived with $N = 84$ examples, and a pairwise dataset P is built. Formally, $P = P_1, ..., P_O$ with $O = c\ \mathrm{C}_2^N = M\ \mathrm{C}_2^{84} = 55776$, where $M = 16$ is the number of EEG channels. The space complexity of the pair dataset is $\mathcal{O}(c\ \mathrm{C}_2^N)$. The SNN training session is done with a batch size multiple of the number of channels. In particular, we use $B = 16 * c$. Therefore, there are 16 pairs of individuals in each batch and each pair of individuals has $c = 16$ channel pairs. This scheme can only be applied in small datasets, since the model does not scale well in terms of space complexity, but our goal is precisely to tackle small datasets with the creation of a whole new optimization space, where the variability contained in the data can be exploited in a different way.

3.3. Validation

Once the SNN has been tuned and trained (in a 20 epochs session), the outputs of the BN for every example were the result of our feature extraction process. With these features, the following classifiers were trained to identify schizophrenia: Support Vector Machines (SVM), Random Forest (RF), XGBoost (XGB), Naive Bayes (NB) and k-Nearest Neighbors (kNN). This process, illustrated in Figure 3, was performed with a LOOCV, where each fold consists on one subject (16 channels/instances). For each of these classifiers, BO hyperparameter tuning is also performed, setup with a maximum of 10 acquisitions and 5 iterations for initial exploration. Algorithm 1 describes the validation schema. The hyperparameter domains for each classifier were:

- SVM: type of kernel (linear or radial-basis function kernel), cost $C \in [0.5, 5]$, and gamma coefficient $\gamma \in [0.00001, 1.0]$
- RF: number of estimators $N_e \in \{5, 10, 15, 20, 25\}$
- XGB: maximum depth $d \in \{3, 4, 5, 6, 7\}$, learning rate $\lambda \in [0.001, 0.1]$, and number of estimators $N_e \in \{10, 50, 100, 200\}$

- NB has no hyperparameters
- kNN: number of neighbors $k \in \{2, 3, 4, 5, 6, 7, 8\}$

Algorithm 1 Leave One Subject Out Cross Validation.

```
predictions ← {}
for each $X_i \in X$ do
    train ← $X \setminus \{X_i\}$
    paired_train ← pair_structure(train)
    SNN.hyperparameters ← SNN.BO(paired_train)
    snn ← SNN.fit(paired_train)
    extracted_features_train ← snn.BN.predict(train)
    extracted_features_$X_i$ ← snn.BN.predict($X_i$)
    classifier.hyperparameters ← classifier.BO(extracted_features_train)
    clf ← classifier.fit(extracted_features_train)
    test_prediction ← clf.predict(extracted_features_$X_i$)
    predictions ← predictions ∪ mean(test_prediction)
end for
return predictions
```

The hyperparameter tuning optimization for the classifiers is also performed in a K-Fold Cross Validation setup ($K = 5$), but instead of using the whole dataset (as was the case for the SNN) only the training set of the LOOCV partition was used. Similar to the BO for the SNN, the combination of hyperparameters with the best average validation accuracy is chosen for each classifier.

4. Results

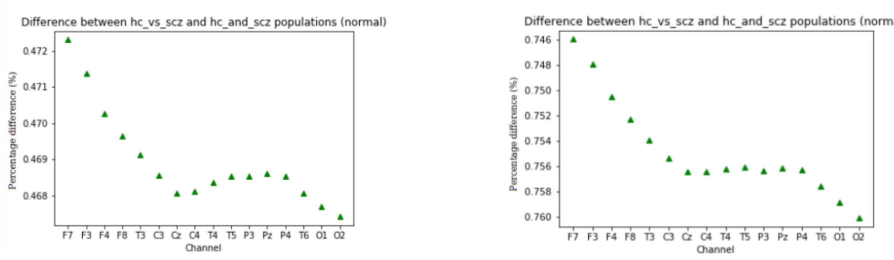
Given the recording setting introduced in Section 3.1 consider the two following sets of paired individuals:

- *hc_vs_scz* - set of pairs of non-neighbor individuals (healthy controls paired with schizophrenic);
- *hc_and_scz* - set of pairs of neighbor individuals (healthy controls paired with healthy controls plus schizophrenic paired with schizophrenic).

Figure 5 shows the spectral differences using FFT between concordant pairs of individuals (*hc_and_scz*) and discordant pairs of individuals (*hc_vs_scz*). Delineate differences would indicate the possibility to correctly group individuals. However, the gathered differences are remarkably low – less than 1% for every channel –, confirming the difficulty of discriminating true pairs of individuals. Despite the nearly absent differences, cosine distance achieves higher percentage differences than the Euclidean distance, motivating its choice for the contrastive loss.

	Classifier	Accuracy	Sensitivity	Specificity	MCC
(i)	FFT-kNN	0.60 ± 0.31	0.56 ± 0.33	0.64 ± 0.30	0.17
(ii)	FFT-NB	0.57 ± 0.32	0.33 ± 0.38	**0.85 ± 0.14**	0.18
(iii)	FFT-RF	0.58 ± 0.32	0.58 ± 0.32	0.64 ± 0.29	0.19
(iv)	FFT-SVM	0.66 ± 0.28	0.69 ± 0.26	0.63 ± 0.29	0.30
(v)	FFT-XGB	0.65 ± 0.28	0.68 ± 0.26	0.61 ± 0.30	0.26
(vi)	EEGNet [8]	0.58 ± 0.32	0.58 ± 0.31	0.59 ± 0.32	0.17
(vii)	EEGNet-SSVEP [8]	0.54 ± 0.34	0.60 ± 0.31	0.46 ± 0.37	0.04
(viii)	Riemann [7]	0.41 ± 0.50	0.47 ± 0.54	0.44 ± 0.50	−0.10
(ix)	DeepConvNet [9]	0.54 ± 0.12	0.64 ± 0.08	0.41 ± 0.14	0.01
(x)	ShallowConvNet [9]	0.57 ± 0.32	0.58 ± 0.31	0.56 ± 0.32	0.12
(xi)	DSTFT-SNN-kNN	0.88 ± 0.12	0.90 ± 0.09	0.85 ± 0.14	0.74
(xii)	DSTFT-SNN-NB	0.83 ± 0.16	0.82 ± 0.16	0.83 ± 0.15	0.62
(xiii)	DSTFT-SNN-RF	0.88 ± 0.11	0.93 ± 0.07	0.82 ± 0.16	0.71
(ixx)	DSTFT-SNN-SVM	0.87 ± 0.12	0.96 ± 0.04	0.78 ± 0.20	0.74
(xx)	**DSTFT-SNN-XGB**	**0.95 ± 0.05**	**0.98 ± 0.02**	**0.92 ± 0.07**	0.88

Table 2: Comparison of classifiers based on discriminative spectral features, state-of-the-art EEG data classifiers, and the proposed SNN-based classifiers. Sensitivity refers to the porpotion of actual Schizophrenic individuals correctly classified. Specificity refers to the proportion of actual Healthy Control individuals correctly classified. Accuracy refers to the proportion of Schizophrenic and Healthy Controls correctly classified. MCC refers to the Matthews Correlation Coefficient [37], which allows us to analyze the significance of our results based on the number of instances.



(a) Euclidean Distance.

(b) Cosine Distance.

Figure 5: Distance type comparison on *hc_and_scz* and *hc_vs_scz* sets.

To assess the proposed contributions, classification results were collected using the extracted features from the developed SNN, and compared with state-of-the-art classifiers developed by Schirrmeister [9], Charles [7] and Lawhern [8]. We further compare our approach against classifiers able to learn directly from spectral/FFT features extracted from each channel [38]. The EEG classifiers proposed in previous works are referred to as: (vi) EEGNet, (vii) EEGNet-SSVEP, (viii) Riemann, (ix) DeepConvNet, (x) ShallowConvNet. The FFT features classifiers are referred to as: (i) FFT-kNN, (ii) FFT-NB, (iii) FFT-RF, (iv) FFT-SVM, (v) FFT-XGB. The proposed classifiers based on the SNN extracted features are referred to as: (xi) DSTFT-SNN-kNN, (xii) DSTFT-SNN-NB, (xiii) DSTFT-SNN-RF, (ixx) DSTFT-SNN-SVM, (xx) DSTFT-SNN-XGB.

According to Table 2, the SNN features outperform the baselines considered by an average of 20pp both in accuracy, specificity and sensitivity. In fact all of the collected differences are statistically significant under significance thresholds below 1E-5. Further, the Matthews Correlation Coefficient (MCC) values for the SNN features show that they are significant. The MCC varies between $[-1, 1]$ and a high value means the results are significant. Figures 6a and 6b show the values of the SNN features for healthy controls and schizophrenic individuals, as well as the statistical significance of each feature.

The results observed when considering FFT features underline the difficult nature of the problem at hands, showing that the use of spectral features is not sufficient to capture discriminative electrophysiological brain patterns.

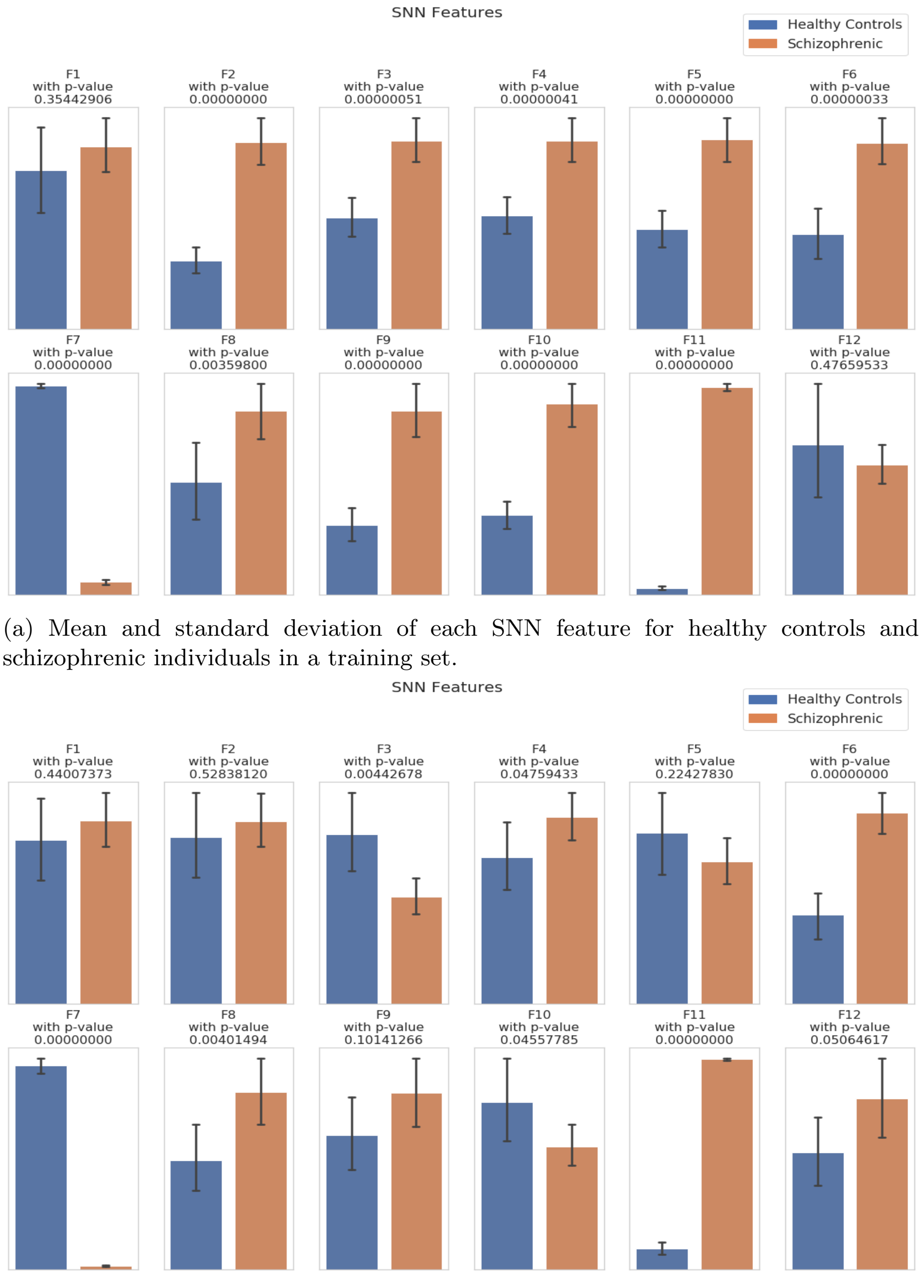
As previously mentioned in Section 2, the previous work on EEG data classification – referred in Table 2 as (vi), (vii), (viii), (ix) and (x) – is unable to capture neuroplasticity differences between healthy and Schizophrenia individuals from resting state data. These approaches are mainly prepared to detect evoked potentials in response to specific stimuli, thus generally neglecting subtle, spontaneous electrophysiological variations in the brain of individuals.

In contrast, the use of DSTFT representations followed by application of the proposed SNNs are better prepared to detect neuroplasticity characteristics on the EEG signal as motivated by the rich spectral content inputted to the SNN, the properties of the entailed transformations, and the discriminative power of the features outputted from the SNN. These observations are experimentally demonstrated by the results presented in Table 2, with a significant difference between our approach and the previous work on EEG.

Among the classifiers applied to the SNN features, XGBoost has the better performance, followed by RFs, SVMs with sparse kernel and kNNs. We hypothesize that this observation is primarily driven by the compositional value of the extracted features and the heterogeneity of individual profiles. Understandably, since only a part of the overall features have discriminative value for a given subject due to profile heterogeneity, NB and kNN have an understandable lower performance due to their inherent inability to discard non-relevant features. Similarly, when we compare the classifiers performance from FFT features, FFT-kNN and FFT-NB have a slightly inferior performance against FFT-XGB and FFT-SVM. Among the five classifiers all of them slightly underperformed on discriminating healthy controls (specificity) than discriminating schizophrenic individuals (sensitivity) due to an inherent ability to avoid false negatives.

Considering an i7-8550U CPU @ 1.88GHz processor with 8GB RAM, the computational complexity of performing a diagnostic (testing a new individual) is residual, below 0.01 seconds. Once hyperparameters are fixed, training the top-performing DSTFT-SNN-XGB classifier from scratch on the target population is achieved in less than 60 seconds.

The gathered results confirm the relevance of working in a pairwise distance space to guarantee a good generalization ability. In addition, the applied con-



(a) Mean and standard deviation of each SNN feature for healthy controls and schizophrenic individuals in a training set.

(b) Mean and standard deviation of each SNN feature for healthy controls and schizophrenic individuals in a test set.

Figure 6: Statistical analysis of the SNN features in a 80/20 train and test setting. The hyperparameters were obtained from a random fold of the LOOCV.

volution transformations guarantee a sensitivity to the inherently rich spatial, temporal and spectral nature of the EEG signal. We hypothesize that these aspects, together with the use of regularization and the cosine loss function (able to favor variations over absolute differences in the spectral content), explain the ability to learn extremely discriminative features.

5. Conclusion

Schizophrenia patients have been associated with deficient neuroplasticity properties [39] present in frequency domain of the EEG signals [40]. Further, this property has been addressed by applying neurofeeback techniques and analysing the corresponding changes in the frequency domain of the EEG signal [1]. With the same goal of extracting this properties, a representation of the EEG signal in the frequency domain, by means of the DSTFT, was processed by a deep learning architecture capable of taking advantage of the properties previously shown by studies related with neuroplasticity in schizophrenia patients.

The rich nature of the electrophysiological data measured at the cerebral cortex makes deep learning a natural candidate to study disorders disrupting the normal brain activity. Nevertheless, the limited size of case-control populations, together with the inherent variability of the spectral content within and among individuals, had left the value of neural network approaches largely unexplored. This manuscript stresses the relevance of revisiting this problem, showing that adequately reshaped neural networks with proper loss and regularization can increase the accuracy of Schizophrenia diagnostics by 15-to-20 percentage points against peer alternatives (without hampering sensitivity or specificity).

Two master principles underlie these results: 1) the mapping of the original data space into a pairwise distance space to support data augmentation while enhancing the discriminative power of the output features; and 2) the exploration of the rich nature of brain patterning through convolution operations on the spectral imaging of the signal, with weights learned under a cosine loss to improve robustness against the inherent noisy nature of electrophysiologic data.

As future work, we aim to extend the experimental analysis towards alternative disorders, and different EEG instrumentation or protocols; contrast the performance of the proposed EEG-based learners against state-of-the-art MRI- and PET-based learners on a population of individuals with (and without) neurodegenerative conditions being currently monitored at Instituto de Medicina Molecular; and to establish a method that is capable of performing a neurofeedback technique to tackle Schizophrenia symptoms, similarly to what has been previously proposed by Nan et al. [1].

[1] W. Nan, J. P. Rodrigues, J. Ma, X. Qu, F. Wan, P.-I. Mak, P. U. Mak, M. I. Vai, A. Rosa, Individual alpha neurofeedback training effect on short term memory, International Journal of Psychophysiology 86 (1) (2012) 83 – 87. doi:https://doi.org/10.1016/j.ijpsycho.2012.07.182.

[2] A. J. Fowle, C. D. Binnie, Uses and abuses of the eeg in epilepsy, Epilepsia 41 (2000) S10–S18.

[3] M. J. Owen, A. Sawa, P. B. Mortensen, Schizophrenia, The Lancet 388 (10039) (2016) 86 – 97. doi:https://doi.org/10.1016/S0140-6736(15)01121-6.
URL http://www.sciencedirect.com/science/article/pii/S0140673615011216

[4] W. Nan, F. Wan, L. Chang, S. H. Pun, M. I. Vai, A. Rosa, An exploratory study of intensive neurofeedback training for schizophrenia, Behavioural neurology 2017.

[5] N. N. Boutros, C. Arfken, S. Galderisi, J. Warrick, G. Pratt, W. Iacono, The status of spectral eeg abnormality as a diagnostic test for schizophrenia, Schizophrenia research 99 (1-3) (2008) 225–237.

[6] B. J. Roach, D. H. Mathalon, Event-related eeg time-frequency analysis: an overview of measures and an analysis of early gamma band phase locking in schizophrenia, Schizophrenia bulletin 34 (5) (2008) 907–926.

[7] P. Charles, pyriemann, https://github.com/alexandrebarachant/pyRiemann, accessed October 22^{nd} 2019 (2013).

[8] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, B. J. Lance, Eegnet: a compact convolutional neural network for eeg-based braincomputer interfaces, Journal of Neural Engineering 15 (5) (2018) 056013.
URL http://stacks.iop.org/1741-2552/15/i=5/a=056013

[9] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, T. Ball, Deep learning with convolutional neural networks for eeg decoding and visualization, Human brain mapping 38 (11) (2017) 5391–5420.

[10] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical image analysis 42 (2017) 60–88.

[11] F. LopesdaSilva, Eeg and meg: Relevance to neuroscience, Neuron 80 (5) (2013) 1112 – 1128. doi:https://doi.org/10.1016/j.neuron.2013.10.017.
URL http://www.sciencedirect.com/science/article/pii/S0896627313009203

[12] K. Gorbachevskaya, S. Borisov, Eeg of healthy adolescents and adolescents with symptoms of schizophrenia, http://brain.bio.msu.ru/eeg_schizophrenia.htm, online; accessed 1st February 2019 (2002).

[13] G. Koch, R. Zemel, R. Salakhutdinov, Siamese neural networks for one-shot image recognition, in: ICML Deep Learning Workshop, Vol. 2, 2015, p. 0.

[14] F. Riaz, A. Hassan, S. Rehman, I. K. Niazi, K. Dremstrup, Emd-based temporal and spectral features for the classification of eeg signals using supervised learning, IEEE Transactions on Neural Systems and Rehabilitation Engineering 24 (1) (2016) 28–35. doi:10.1109/TNSRE.2015.2441835.

[15] P. Zhang, X. Wang, W. Zhang, J. Chen, Learning spatialspectraltemporal eeg features with recurrent 3d convolutional neural networks for cross-task mental workload assessment, IEEE Transactions on Neural Systems and Rehabilitation Engineering 27 (1) (2019) 31–42. doi:10.1109/TNSRE.2018.2884641.

[16] S. Fazli, M. Danczy, J. Schelldorfer, K.-R. Mller, L1-penalized linear mixed-effects models for high dimensional data with application to bci, NeuroImage 56 (4) (2011) 2100 – 2108. doi:https://doi.org/10.1016/j.neuroimage.2011.03.061.
URL http://www.sciencedirect.com/science/article/pii/S1053811911003405

[17] A. Destrero, S. Mosci, C. De Mol, A. Verri, F. Odone, Feature selection for high-dimensional data, Computational management science 6 (1) (2009) 25–40.

[18] J. Snoek, H. Larochelle, R. P. Adams, Practical bayesian optimization of machine learning algorithms, in: Advances in neural information processing systems, 2012, pp. 2951–2959.

[19] A. Bhandari, D. Voineskos, Z. J. Daskalakis, T. K. Rajji, D. M. Blumberger, A review of impaired neuroplasticity in schizophrenia investigated with non-invasive brain stimulation, Frontiers in psychiatry 7 (2016) 45.

[20] K. K. Liu, R. P. Bartsch, A. Lin, R. N. Mantegna, P. C. Ivanov, Plasticity of brain wave network interactions and evolution across physiologic states, Frontiers in neural circuits 9 (2015) 62.

[21] T. Shaikhina, N. A. Khovanova, Handling limited datasets with neural networks in medical applications: A small-data approach, Artificial Intelligence in Medicine 75 (2017) 51 – 63. doi:https://doi.org/10.1016/j.artmed.2016.12.003.
URL http://www.sciencedirect.com/science/article/pii/S0933365716301749

[22] B. Barz, J. Denzler, Deep learning on small datasets without pre-training using cosine loss (2019). arXiv:1901.09054.

[23] G. Hu, X. Peng, Y. Yang, T. M. Hospedales, J. Verbeek, Frankenstein: Learning deep face representations using small data, IEEE Transactions on Image Processing 27 (1) (2017) 293–303.

[24] Z. Dvey-Aharon, N. Fogelson, A. Peled, N. Intrator, Schizophrenia detection and classification by advanced analysis of eeg recordings using a single electrode approach, PloS one 10 (4) (2015) e0123033.

[25] Z. Dvey-Aharon, N. Fogelson, A. Peled, N. Intrator, Connectivity maps based analysis of eeg for the advanced diagnosis of schizophrenia attributes, PloS one 12 (10) (2017) e0185852.

[26] M. Sabeti, S. Katebi, R. Boostani, Entropy and complexity measures for EEG signal classification of schizophrenic and control participants, Artificial Intelligence in Medicine 47 (3) (2009) 263–274. doi:10.1016/j.artmed.2009.03.003.
URL https://doi.org/10.1016/j.artmed.2009.03.003

[27] S. Sengupta, A. Singh, H. A. Leopold, T. Gulati, V. Lakshminarayanan, Ophthalmic diagnosis using deep learning with fundus images a critical review, Artificial Intelligence in Medicine 102 (2020) 101758. doi:https://doi.org/10.1016/j.artmed.2019.101758.
URL http://www.sciencedirect.com/science/article/pii/S0933365719305858

[28] C. Ieracitano, N. Mammone, A. Bramanti, A. Hussain, F. C. Morabito, A convolutional neural network approach for classification of dementia stages based on 2d-spectral representation of eeg recordings, Neurocomputing 323 (2019) 96 – 107. doi:https://doi.org/10.1016/j.neucom.2018.09.071.
URL http://www.sciencedirect.com/science/article/pii/S0925231218311524

[29] S. L. Oh, J. Vicnesh, E. J. Ciaccio, R. Yuvaraj, U. R. Acharya, Deep convolutional neural network model for automated diagnosis of schizophrenia using eeg signals, Applied Sciences 9 (14) (2019) 2870.

[30] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, R. Shah, Signature verification using a" siamese" time delay neural network, in: Advances in neural information processing systems, 1994, pp. 737–744.

[31] A. B. Hill, Principles of medical statistics, The Lancet, 1955.

[32] F. M. Howells, H. S. Temmingh, J. H. Hsieh, A. V. van Dijen, D. S. Baldwin, D. J. Stein, Electroencephalographic delta/alpha frequency activity differentiates psychotic disorders: a study of schizophrenia, bipolar disorder and methamphetamine-induced psychotic disorder, Translational psychiatry 8 (1) (2018) 75.

[33] R. H. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, H. S. Seung, Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit, Nature 405 (6789) (2000) 947.

[34] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), Vol. 2, IEEE, 2006, pp. 1735–1742.

[35] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, Journal of Machine Learning Research 15 (2014) 1929–1958.
URL `http://jmlr.org/papers/v15/srivastava14a.html`

[36] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.

[37] B. Matthews, Comparison of the predicted and observed secondary structure of t4 phage lysozyme, Biochimica et Biophysica Acta (BBA) - Protein Structure 405 (2) (1975) 442 – 451. `doi:https://doi.org/10.1016/0005-2795(75)90109-9`.
URL `http://www.sciencedirect.com/science/article/pii/0005279575901099`

[38] H. Hindarto, S. Sumarno, Feature extraction of electroencephalography signals using fast fourier transform, CommIT (Communication and Information Technology) Journal 10 (2) (2016) 49–52.

[39] J. K. Wynn, B. J. Roach, A. McCleery, S. R. Marder, D. H. Mathalon, M. F. Green, Evaluating visual neuroplasticity with eeg in schizophrenia outpatients, Schizophrenia Research 212 (2019) 40 – 46. `doi:https://doi.org/10.1016/j.schres.2019.08.015`.
URL `http://www.sciencedirect.com/science/article/pii/S0920996419303561`

[40] R. Zomorrodi, T. Rajji, D. Blumberger, Z. Daskalakis, The association between cross-frequency coupling and neuroplasticity via paired associative stimulation: Tms-eeg study, Brain Stimulation 12 (2) (2019) 512. `doi:https://doi.org/10.1016/j.brs.2018.12.680`.
URL `http://www.sciencedirect.com/science/article/pii/S1935861X18310994`