CLASSIFICATION AND ANALYSIS OF EEG SIGNALS FOR BRAIN-COMPUTER INTERFACE

Student: Anna Grillenberger (s213637); Supervisors: Silvia Tolu, Sadasivan Puthusserypady

ABSTRACT

Brain Computer-Interfaces (BCI) proved to be an effective method for the rehabilitation of post-stroke impairments by translating brain signals into movement intentions and supporting the execution of the motion with external devices such as exoskeletons. This is one of many applications through which the classification of electroencephalography (EEG) recordings of motor imagery (MI) tasks gained a lot of attention in current research. Different methods have been used to extract distinctive features from the recordings, with Deep Learning (DL) models achieving the best results. However, training DL models requires a lot of data to result in a robust and accurate performance. This work presents an approach to how DL models can be trained to classify EEG-MI recordings with only a few data available. By making use of the public dataset PhysioNet, a base model is trained with data of many subjects, learning to extract features from EEG data. This pre-learned knowledge is exploited when fine-tuning the model on one specific subject with little data from our recorded dataset. After fine-tuning the PhysioNet base model on 11 individual subjects and selecting the most robust channel pairs, a median accuracy of 98% is reached. Overall, the approach of fine-tuning with a different dataset offers new insights for the development of individualized real-world BCI applications where limited data is available.

***Index Terms*—** BCI, EEG-MI, neural networks, transfer learning, small dataset

1. INTRODUCTION

Stroke is the second leading cause of death [1] and one of the main causes of long-term disability like motor impairments or limb paralysis [2]. However, with effective rehabilitation of physical therapy, stroke patients could partially regain their motor control [3]. Research has demonstrated that BCI systems are an effective method for motor neurorehabilitation of patients with post-stroke impairments as it serves as a communication channel between the brain and the outside world. This is achieved through the ability of BCI to bypass the normal motor output pathways by directly decoding and translating brain electrophysiological signals into motor intentions that can be used to control robotic systems [4]. It is exploited that imagining a movement, called Motor Imagery (MI), activates the same area in the brain (motor cortex) as when the movement is actually performed [5]. Therefore, BCI systems can close the loop between translating MI recordings into motor intentions and controlling a robot to support the respective body part in performing the passive movement as imagined.

To record human intentions with brain activity, non-invasive techniques based on Electroencephalogram (EEG) are easy-to-use and less risky as they don't require surgery. EEG reflects the brain activity caused by electrical charges of different neuron populations in the central nervous system, measured by electrodes. Recently, the utilization of both commercial and free-scale EEG devices is expanding [6], making them cost-effective tools suitable for future commercial BCI systems [7].

However, EEG signals have a low spatial resolution, a high temporal resolution, and a low signal-to-noise ratio (SNR) [8]. Additionally, high differences between different subjects and between different trials of one subject can be observed, leading to a large variability in the signal. These variations can be caused by artifacts such as blinking, muscle movement, or electrical interference recorded by the EEG systems, as well as changes in mood, fatigue, or attention of the subject [9]. These signal characteristics are challenging and lead to poorer classification results. Therefore, the development of methods to effectively preprocess the noisy data, extract features, and classify the EEG signals has become an important research topic [10].

Traditional methods for MI-EEG classification usually address preprocessing, feature extraction, and classification separately. After bandpass filtering the signal to the expected bandwidths, feature extraction methods as variations of Common Spatial Pattern (CSP) [11], Fast Fourier Transform (FFT) [12], or Wavelet Transform [12], are used to reduce the dimensions of the data while keeping the most important attributes. Supervised machine learning methods like Support Vector Machine (SVM) [13], Linear Discriminant Analysis (LDA) [14], or k-nearest-neighbor (KNN) [15] are applied to classify the input instance based on the extracted key features. However, valuable information from EEG signals may be lost during the feature extraction process.

In contrast to that, DL methods have attracted attention

offering a distinct advantage over traditional methods as they provide an end-to-end solution, leading to superior performances. As demonstrated in the work of Tibrewal et al., comparing a classical approach of CSP and Linear Discriminant Analysis with a neural network, the DL model improved the classification accuracy for all subjects by 28.28% [16]. As DL models can effectively deal with nonlinear and non-stationary data, it is a suitable solution to learn underlying high-level and latent complex features from signals. With preprocessed EEG signals as input, DL models use their deep architecture to automatically extract features and classify the data in a single model, streamlining the entire process [8]. In recent years, several novel DL approaches have been proposed for EEG-based BCI [8]. Many of them are based on convolutional neural networks (CNN) due to their ability to learn abstract features from local receptive fields. Using smaller filters than the input pattern allows to focus on spatially local features (sparse interaction). As the weights of the filter are trained and the same filter is applied to the whole input data, parameters are shared, leading to a drastic decrease in the number of needed parameters and computational efficiency. Moreover, low-level features can be extracted by processing the data with convolutional layer repetitions, making preprocessing unnecessary. This makes CNNs a suitable approach for complex EEG recognition tasks, achieving good results and being widely used for EEG-MI classification [17], [18].

However, in comparison to classical methods like CSP, deep neural networks require a large amount of data in order to train the numerous parameters in all layers. Since collecting EEG-MI data is time-consuming and mentally exhausts the subject after some time of recording, affecting the quality of data, the amount of data per subject is small. To address the issue of limited subject-specific data, subject-independent approaches have been researched more recently, where a model is trained on data of multiple subjects [19]. With this approach, the challenge of finding a common model for data of high inter-subject variability becomes even more prominent. Therefore, the method of fine-tuning a subject-independent model on subject-specific data can be used to achieve an adaption to the target subject, leading to superior classification results [20].

This work presents an approach to how a small dataset of EEG-MI data for stroke rehabilitation can be used to train a DL model for the classification of left and right arm movements. As the training of DL models requires a lot of data, this work elaborates on how a bigger publicly available BCI dataset can be used to address the lack of data, pointing out the differences in measurement setting and MI-task. Different fine-tuning approaches and data augmentation methods for overcoming the challenges of limited data and variety in EEG-MI data are evaluated based on the resulting classification accuracies.

The main contributions of this work are as follows: First, we will explore the capabilities of a model, trained on more than 100 subjects of the PhysioNet dataset, to classify new subjects. This subject-independent performance gives insights if a trained base model can be applied to different, smaller datasets. Next, a fine-tuning approach is implemented and evaluated, giving insights on the effectiveness of training a base model beforehand, which channel pairs should be selected, and which data augmentation method leads to the highest classification accuracies.

2. RELATED WORK

In the last years, the use of DL methods for MI classification has increased rapidly [21]. Next to CNNs, many other architectures have been proposed in research, such as Recurrent Neural Networks (RNN), Auto-encoders (AE), or Deep Belief Networks (DBN). Some studies used RNNs or Long short-term memory (LSTM) to extract temporal information from the EEG signals [22], [23]. Overall, the performance of CNN models has shown superior results in MI classification tasks compared to other architectures [21]. Therefore, CNNs have been fused with other DL models as Multi-layer perceptron or Auto-encoder architectures where the fusion methods outperformed all the state-of-the-art machine learning and DL techniques for EEG classification at that time [24].

A drawback of many of the mentioned DL studies is, that combined EEGs from multiple trials per person are utilized to train subject-specific models. However, one of the biggest challenges in BCI projects with DL is the lack of subject-specific data to train the networks effectively. Therefore, more recently, work has been proposed in literature where pre-existing data not only from one, but from multiple other subjects is used to train a model, often exploiting the availability of large public datasets such as PhysioNet [25] or BCI Competition IV [26]. However, transferring knowledge from other subjects is challenging due to high inter-subject variabilities [27]. Therefore, many approaches have restricted adaptability and may lack robustness when applied to different individuals.

To address this issue, cross-individual validation comes into focus where the model is tested on new, unseen subjects. Even though this training technique is more challenging as it requires information transfer between different individuals, the models evaluated by this approach were more robust and generalized [24]. Recently, research is combining attention mechanisms with DL models where based on human brain behavior, the model selectively focuses on a few significant elements while others are ignored. With these transformer architectures, mostly known from language models like BERT [28] and GPT-2 [29], long-range dependencies can be analyzed without using complex recurrent or CNNs [30] while having better interpretability than other DL models [31]. Xie et al. designed Transformer-based models for two-, three-, and four-class MI-EEG classifications based on the PhysioNet

dataset, outperforming other state-of-the-art models in cross-individual validation [19]. Similar results were achieved by Altaheri et al., using a multi-head self-attention to highlight the most valuable features and extracting high-level temporal features with a temporal convolutional network (TCN) [32].

Although with these Transformer networks, good results with accuracies around 83% on a two-class MI-task were achieved, the training is nevertheless dependent on the requirement that data of many different subjects is available with the constraint that the settings of data acquisition should be as similar as possible for the model to be able to generalize and transfer the knowledge well to new subjects. However, at the beginning of the development of a BCI system, as in our case, a large amount of recorded measurements is not available. Therefore, another approach to address the difficulties of high inter-subject variabilities is to adapt a pre-existing model on a target subject by fine-tuning it on few subject-specific data.

Dose et al. built a unified end-to-end model using CNNs for learning features and dimension reduction where a conventional fully connected layer was used for classification. With transfer learning, the global classifier was adapted to single individuals improving the overall mean accuracy from 80.38% to 86.49% for a two-class problem [18].

The work of Zhang et al. studied 5 adaption schemes of a deep CNN with limited EEG data where as well the effect of different learning rates and percentages of adaptation data was evaluated [20]. With a deep CNN network architecture by Schirrmeister et al. as a baseline [17], their aim was to keep the extracted features from the convolution filters of the model while adapting the last classification layers to a new subject with fine-tuning. With a subject-independent MI-classification accuracy of 84.19%, an increase of 3.21% in average accuracy could be achieved with their proposed adaption methodology.

In comparison to using all EEG channels, Mattioli et al. proposed a 10-layer one-dimensional CNN (1D-CNN) where only a limited number of EEG channels was used, favoring cheap applications in the future [33]. Their presented transfer learning method used the EEG group dataset to extract critical features to then customize the model to the single individual by training its late layers with only 12-min individual-related data. In addition, their use of the baseline class is different from previous work with the goal of discriminating between movement intentions and non-relevant or noisy signals. This feature is beneficial for real-life applications.

For this work, the 1D-CNN architecture is used because good results were obtained despite the simplicity of the model. It is expected that a less complex model might be less prone to overfitting with regard to the scarcity of data of our BCI solution. Additionally, by having only 2 channels as input, multiple channel pairs of the measurement can be used, leading to more data available.

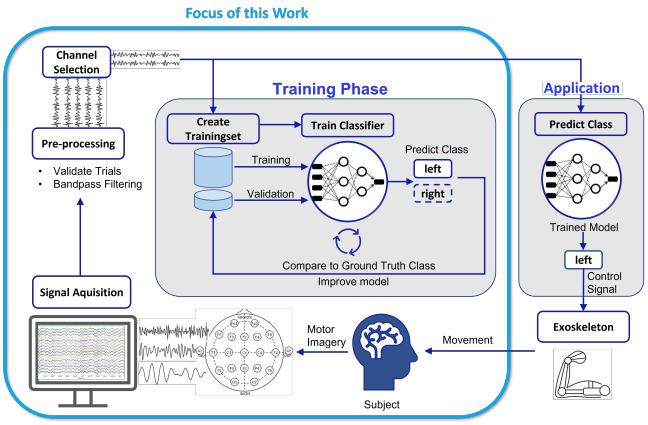


Fig. 1: Overview of a BCI-system for upper limb stroke rehabilitation.

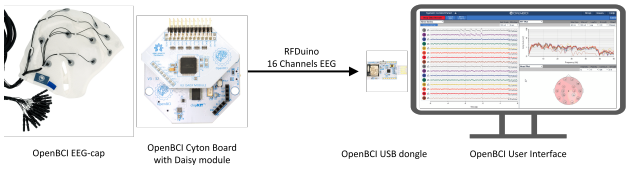


Fig. 2: OpenBCI recording setup.

3. MATERIALS AND METHODS

3.1. System Overview

This work focuses on the development of a classification algorithm to translate MI signals into movement intentions. However, the classifier is only one part of the BCI setup for upper limb stroke patient rehabilitation. The overview of the whole system can be seen in figure 1 where two different phases are described. In the training phase, subjects are asked to perform MI tasks in a controlled measurement environment where the EEG signals can be linked to the corresponding motor intention. This data is pre-processed and used to train a classification model. In the application phase, an already trained classifier is used for rehabilitation. The brain activity of the subject is measured, pre-processed, and passed to the classification model. The model predicts which arm movement (left/right) the subject intended to do and serves as a control signal for the exoskeleton. The exoskeleton will perform the suggested task and help the subject to connect the intention with the actual muscle movement.

3.2. Dataset and Preprocessing

3.2.1. Own Recordings

Recording Setup For recording EEG-MI-data, the dry EEG electrodes cap from Open BCI is used, measuring 16 chan-

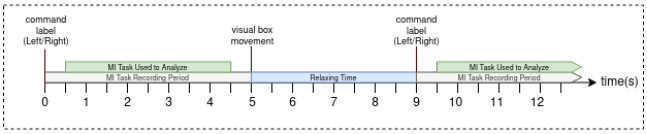
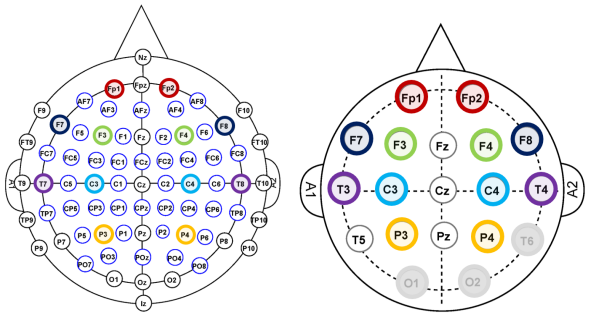


Fig. 3: Measurement protocol of training data acquisition.

	PhysioNet	Own Recordings
Frequency	160 Hz	125 Hz
Channels	64	16
Classes	$3 + baseline$	2
Subjects	109	11
Trials per Subject	42	80
Trial Duration	4 seconds	5 seconds

Table 1: Comparison of dataset settings.



(a) PhysioNet Channels (b) Own channel Setup

Fig. 4: Channel positions of PhysioNet system (a) and our own system (b). The channel pairs in the same color have matching positions in both settings. The greyed-out channels were not used.

nels with 125 Hz. As the dry electrodes are not attached to the skin using a conductive feel, the electrode impedance is higher, leading to more sensitivity regarding movement and lower signal quality. As visualized in figure 2, the EEG cap is connected to the Cyton OpenBCI Board with the Daisy module. It communicates wirelessly to a USB dongle that is connected to the computer. The measurements were organized in one session per recorded subject. All subjects, 7 male and 4 female are healthy with an average age of 25. Each recording session contains 3 measurements with small breaks in between where each measurement contains 20 trials per class. In order to introduce the subjects to the imagery task, the first measurement included slow curl movements of the respective arm. In this way, errors in understanding the recording principle could be easily detected and subjects could prepare mentally for the task ahead. The following two measurements were performed without movement to record the correct MI-task.

Recording Process For gathering the data, a Python script with a visual interface was used. Based on the measurement protocol of figure 3, commands of the randomly selected MI-task are displayed for 5 seconds while recording the data. After that, a relaxation command gives notice about the 4 seconds break. To inform the user about the progress of the measurements, two boxes, corresponding to the left and right MI-tasks, are moving upwards for every accomplished measurement. Starting at the bottom at the beginning, the right box moves upwards if the right MI task is finished and vice versa. Additionally, a validation of the recorded data is implemented. If the size of expected samples exceeds a deviation of $\pm 1\%$ because data packages were lost or the communication was not reliable, the recorded trial is discarded.

Preprocessing After recording the raw data, each trial is checked for validity by removing all trials with values out of the expected range of $\pm 250 \mu V$. After that, a band pass filter between 8 and 15 Hz is applied. Based on Das et al., this sub-band is recommended for classifying MI tasks with good accuracy [34]. The reason for that are the so-called mu rhythms repeating in this selected frequency range of 7.5-12.5 Hz. These mu waves are found in the motor cortex controlling voluntary movement.

3.2.2. Public Dataset - PhysioNet

As the training of a DL model requires a large dataset and the number of our own recordings is limited, the public PhysioNet EEG Motor Movement/Imagery Dataset [25] is used additionally. It consists of 109 subjects with more than 1500 trials. The trials were recorded using the BCI2000 system with 64 electrodes and a sampling rate of 160 Hz. Unlike our recordings, the MI task of PhysioNet is not to flex the upper limb but to open/close the left or right fist. Moreover, a baseline class of rest-state with opened eyes and a MI task between both fists and both feet is available. However, only the two-class trials of L/R MI are used in this work.

Adaption to Our Recording Settings Due to the differences in the settings of the public dataset and our recordings, as seen in table 1, the data of PhysioNet is adjusted to be more similar to our own measurements. The expectation is that the more similar the datasets are, the better the knowledge of the public dataset model can be applied to our data. To adapt the PhysioNet data to our recording settings, the data is resampled from 160 Hz to 125 Hz. In addition, figure 4 shows the different recorded channel locations. Using the PhysioNet data, only 16 out of 64 EEG channels can be used to match our channel locations. Moreover, PhysioNet includes more classes than required in our use case. Therefore, only the

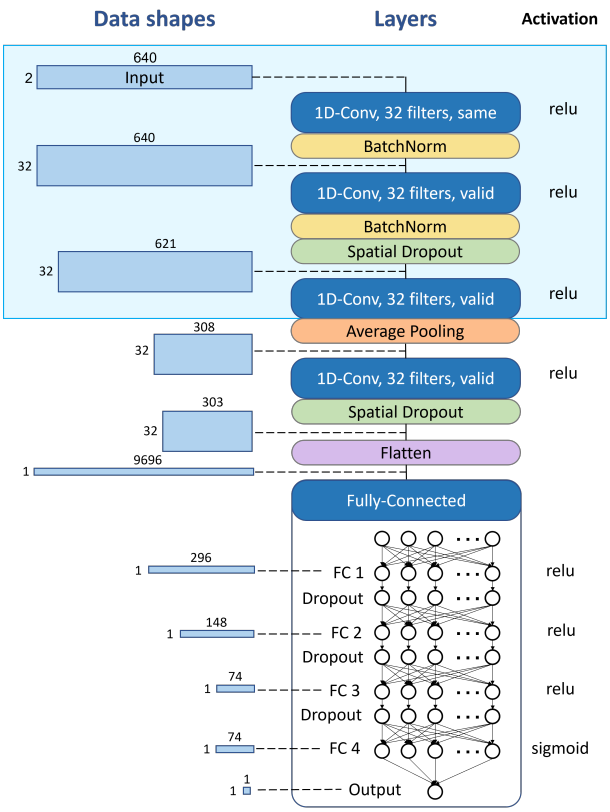


Fig. 5: Architecture of the 1D-CNN model. Layers in the blue rectangle are frozen during fine-tuning.

PhysioNet trials belonging to the left fist/right fist task are selected (experimental runs: 4,8,12). To have the same recording duration, the middle 4 seconds of our 5-second trials were cropped out.

3.3. Model Architecture

The used model architecture is the 1D-CNN of Mattioli et al. [33]. As visualized in figure 5, the architecture consists of four 1D-convolutional layers performing the main task of feature extraction and 4 fully connected layers for classification. The input of the model is a matrix with dimensions M×N where M is the length of the time window (4s * 125Hz = 640) and N is the number of EEG channels which in our case corresponds to two symmetrical channels. This limited number of channels as input enables more simple BCI applications relying on few-channel portable recording devices. Using 1D-CNN layers, the CNN kernels only slide over the time dimension of the input pattern during convolution. The kernels have dimensions of QxN where N always corresponds to the number of channels of the input data and Q is the time window that is covered for one convolution. Using different filters, convolution is performed several times to extract different features. Which features are extracted depends on the filter weights, which in turn are optimized in the training of the model. The formula of the convolution of filter r is the following:

$$y_r = f\left(\sum_{q=1}^{Q}\sum_{n=1}^{N} w_{qn}x_{r+q,r+n} + b\right) \tag{1}$$

The output y_r corresponds to unit r of the filter feature map. The overlapping area of the filter and the two-dimensional input x is multiplied by the filter weight w and a bias b is added. The activation function f adds non-linearity to the filter. The resulting size of the filter feature map is defined by the stride and padding settings. If the stride is 1, the position of the filter is shifted by 1 after each convolution, leading to the same output size as the input signal, when padding is used. Padding adds zero values to the signal/image sides to contain the original dimensions. Therefore, a stride of 1 and padding is called 'same' whereas when padding is not applied, the description 'valid' is used.

In figure 5, the layers of the model, as well as the respective data shapes, are explained. The first convolutional layer uses 32 filters of size 20 to extract low-level features of the time series while preserving the input size with padding. After that, batch normalization is applied before passing the data to the next layer. This usually leads to an increased learning speed and improves the generalization of the network preventing overfitting. After layers 4 and 5, a spatial dropout is applied. By discarding entire feature maps with a probability of 50%, the network is forced to learn redundant representations of the input, making it more robust and less likely to overfit the training data. The average pooling layer reduces the parameter size and therefore also the computation of the network by averaging the feature map values in a neighborhood of size 2x1. Therefore, with a stride of 2, the size is reduced by half. In addition to the reduction of the parameters, the pooling layer makes the representation space more invariant for small input deviations. In the last four fully connected layers, the extracted features are classified into the output class. In each layer l, all neurons are connected to all I units from the previous layer, being computed as

$$y_j^{(l)} = f\left(\sum_{i=1}^{I} w_{ji}^{(l)} \cdot x_i^{(l-1)} + b_j^{(l)}\right) \tag{2}$$

with $w_{ji}(l)$ as the weight of the connection between unit j of this layer and unit i of the previous layer and $b_j(l)$ as the bias of unit j. The nonlinear activation function f of the hidden layers (ReLU) sets all negative input values to zero. As in our MI-classification problem, only 2 classes are predicted, instead of the softmax function from the original architecture, the sigmoid activation with output values between 0 and 1 is used. With a threshold of 0.5, lower values can be classified as class 1 and higher values as class 2.

3.4. Fine-Tuning

Fine-tuning in DL allows to leverage the knowledge of a pre-trained model and adapt it to a new task or dataset. Usually, the pre-trained model is trained on a large dataset where it learned to extract general features from the data. Making use of this knowledge, adapting a pre-trained model to new data requires fewer data than training a new model from scratch.

The process of fine-tuning involves freezing some layers of the pre-trained model and eventually adding new layers on top of it. If a layer is frozen, it means that its parameters are not trainable, so the weights and biases are not updated during training. In this way, general knowledge is preserved while task-specific features are learned by the trainable layers.

The blue rectangle of figure 5 shows which layers of our architecture are frozen in the fine-tuning process. Therefore, the first three convolutional layers are fixed for feature extraction while the last convolutional layer and the fully-connected layers are adapted to classify subject-specific data.

3.5. Data Augmentation

Data augmentation is a technique used in DL to artificially increase the amount of training data by creating new data samples from the existing data. Random transformations are applied to the original data with the goal of creating similar samples that increase the diversity of the training data. Higher variability of the training data helps the model to generalize better to unseen data and reduces overfitting. Moreover, the lack of available data is a problem that often occurs when training DL models which require large amounts of training samples to learn underlying patterns and features of the data. Data augmentation helps to increase the effective size of the training data. Since it is very time-consuming to record MI-BCI datasets and therefore only little data is available, data augmentation comes into focus in this work.

The Synthetic Minority Over-sampling Technique SMOTE is a data augmentation method used in machine learning to address the problem of class imbalance. It selects one or more of its k nearest neighbors and creates new samples by interpolating between them. In this work, this method is not used to compensate for an unequal class ratio but to duplicate samples of both classes.

Surrogating is another data augmentation technique used for time series data that involves creating new, synthetic time series while preserving the statistical properties of the original sample. Firstly, the Fourier transform of the original time series is computed. Then, a set of surrogate time series is generated by randomly shuffling the phases of the Fourier coefficients while keeping their magnitudes fixed. In order to obtain the synthetic time series in the time domain, the Fourier transform of the surrogate time series is inversed. The resulting augmented sample is similar to the original time series in terms of statistical properties, such as mean, variance, and autocorrelation, but is different in its time-domain behavior.

In addition to augmenting the data to obtain more training data, the recorded MI tasks with movement are added to the training set. In this way, more data is available to learn patterns from, but the resulting model performance is not distorted as these recordings with movement are not part of the testing set.

4. EXPERIMENTAL SETUP

4.1. Training Parameter Settings

For optimizing the network parameters, the binary cross-entropy loss function and the Adam optimizer are used to update the CNN parameters. With the Adam optimization, the learning rate adapts based on the moments of the gradient. The hyperparameters are set as follows: $\alpha = 10^{-4}$, $\beta 1 = 0.9$, $\beta 2 = 0.999$, $\epsilon = 10^{-8}$. A batch size of 10 is used for the training of all models. In training the base models, a maximum number of 100 epochs is set. To prevent overfitting and save computational resources, early stopping is used where the training is stopped as soon as the loss of the validation set doesn't improve with a minimum delta of 0.001 for four consecutive epochs. For the fine-tuning process, the training epochs are set to 20, also making use of early stopping. All model training is performed on DTU HPC Tesla V100 with 16GB of RAM. As only few data is used for adapting the model to a specific subject, the training time of the fine-tuned network is significantly lower (approx. 50 times faster) than training the base model.

For the following experiments, the data augmentation method SMOTE is used, increasing the training set by factor 4. Based on the work of Mattioli et al. [33], the best results were achieved with the symmetrical channel pairs [FC1,FC2],[FC3,C4],[C3,C4],[C1, C2],[CP1, CP2], and [CP3, CP4]. Since not all of these channels are available in our BCI setting, the closest ones ([C3,C4],[F3,F4],[P3,P4]) were chosen as an initial channel-pair selection.

4.2. Model Training Techniques

4.2.1. Base-Model Training

As in our BCI-MI recordings, only small data is available, we want to train a base model on a bigger dataset learning to extract features from EEG data. This base model will be evaluated on how well it performs on data it has never seen with the following research questions:

- How does the amount of data influence the training of a subject-independent model? Comparing the ability to

generalize of PhysioNet model (large dataset) and the model trained with our recordings (small dataset)

- How robust is the subject-independent model trained on PhysioNet? Is the performance sufficient when classifying data of new subjects of our dataset?
- With the expectation to perform worse on a different dataset - How big is the difference in accuracy?

For training the base model, the network is evaluated on its ability to generalize to new subjects. This subject-independent performance is evaluated with the Leave-One-Subject-Out cross-validation (CV). This means, that for each test subject, a model was trained on all other subjects. As a consequence, the number of CV folds equals the number of test subjects.

PhysioNet Base-Model In the first experiment, a base model will be trained on the public PhysioNet dataset, leaving out the data of five test subjects (34, 10, 65, 90, 101) from training. The data of these test subjects will be used later to evaluate the fine-tuning performance and the generalizing of the base model. The resulting dataset was split in 80% training and 20% validation data.

Own Dataset Base-Model A second base model, trained on our own small dataset is evaluated. To have as much data available for training as possible, for each target subject, an individual base model was trained on the data of all subjects except the target one. Again, the resulting dataset was split into 80% training and 20% validation data.

4.2.2. Fine-Tuning Methods

To increase the accuracies, the subject-independent base model will be fine-tuned with little training time on the subject-specific data, adapting to each subject individually. For adapting the model, the non-frozen layers are re-trained for more individual classification. For evaluation, a 5-fold cross-validation is implemented. During training, the data of the target subject is randomly split into 5 subsets. One of the 5 subsets is used for testing (20%) and the remaining 4 for training (80%). This process is repeated 5 times until every subset was used once for testing. The resulting 5 accuracies are averaged and provide a more representative model performance than only evaluating the data one time where by chance an above-average good/bad set of test samples could have been chosen. As our classification problem is class-balanced, the chosen performance metric to evaluate the model is the classification accuracy, being the ratio of correctly recognized test samples to all test samples.

Figure 6 visualizes the fine-tuning training. Different fine-tuning methods were evaluated on their classification performance, giving answers to the following questions:

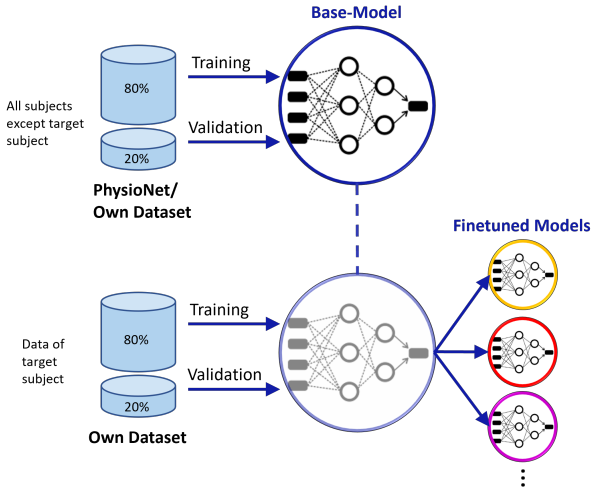


Fig. 6: Training approach of base-model and fine-tuned model.

- Which base model leads to better results when fine-tuning it on our dataset? PhysioNet-Basemodel/Own Data-Base model
- Which channel-pairs should be chosen?
- How do different data augmentation methods influence the accuracy?

In general, subject-independent approaches achieve lower accuracies as the variations of the EEG data between different subjects are significantly higher than within a subject. Therefore, training one general model with a good performance on all subjects poses a greater challenge than a subject-specific solution.

5. EXPERIMENTAL RESULTS

5.1. Classical Methods

As a first step, a traditional approach was used to classify the recorded EEG-MI signals of multiple subjects. After preprocessing the data as described in chapter 3.2.1, the CSP method was used to extract features and reduce the dimensions of the data. It is the most popular algorithm in the BCI field, learning a set of spatial filters where the variance for one class is maximal and minimal for the other one. Then, an SVM model is applied to classify the four best CSP components.

In table 2, the subject-independent and subject-specific accuracy of this approach was evaluated. The training accuracy indicates how well the CSP is able to create distinguishable components from the EEG-data patterns of both classes and with which accuracy the SVM separates these features. However, a high training accuracy does not necessarily mean that the network will perform well on new, unseen data. The

Evaluation Method	Train Accuracy	Test Accuracy
Subject-Independent	70.20%	59.44%
Subject-Specific	83.47%	71.78%

Table 2: Evaluation of CSP and SVM approach. For the subject-specific method, a 10-fold CV was used (90% training, 10% testing). Accuracy is averaged over all 11 test subjects.

ability to generalize to new data is seen in the test accuracy. It can be stated, that the extracted CSP features from 10 subjects can't sufficiently be applied to a new subject, reaching an average accuracy of 59.44%. Therefore, a subject-independent approach using CSP and SVM is not recommended. When training on data of only one subject, and therefore having less variability, an average test accuracy of 71.78% is reached. This supports the fact that data within a subject is more similar and easier to classify. However, an accuracy of around 70% might not be practical for the real-world application of stroke rehabilitation. If too many errors occur, the patient might get frustrated, weakening the acceptance level of BCI systems in real-world applications. Moreover, the effectiveness of the rehabilitation method might be decreased. Therefore, the following sections will elaborate on a DL method where higher accuracies are expected.

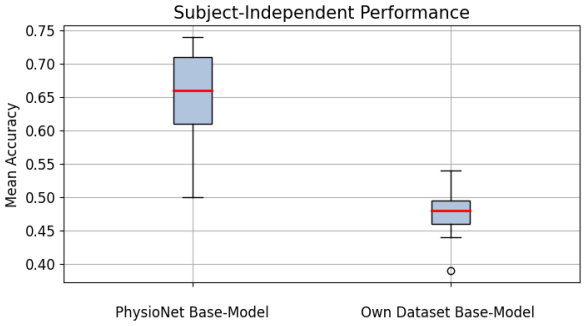
5.2. Deep Learning Approach

5.2.1. Subject-Independent Evaluation

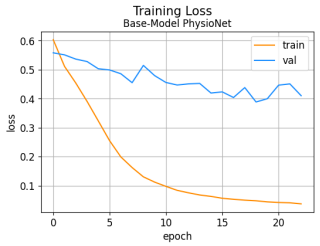
In the first experiment, two different subject-independent models were evaluated in classifying unseen subjects of their dataset. The average classification accuracies of all test subjects are displayed in figure 7a.

The PhysioNet model, trained on 104 different subjects, achieves a median accuracy of 66%. In comparison, with the model trained on our dataset with only 10 training subjects, an accuracy of 48% is reached. Moreover, the training curve of figure 7b shows that during training with the PhysioNet dataset, the network achieves a learning effect as the validation curve declines. However, the overfitting difference between training and validation loss is still large. The reason for that is supposedly the large variabilities between subjects, making it hard to generalize. This overfitting effect is stronger the fewer data is available, which can be observed in figure 7c. Since there is no decline of the validation loss, no learning effect can be achieved when training with few data from our own dataset. This result suggests that it is crucial to how large the dataset is to train a subject-independent model. If only a few data is available, as in our recordings, no satisfactory generalization can be achieved.

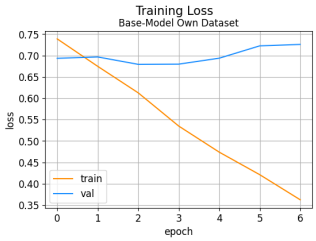
As the PhysioNet model performs better in classifying unseen data of its own dataset, can this knowledge be applied to our data as well? The yellow boxplot in the middle of figure 8



(a) Comparison of subject-independent accuracy of base-models.



(b) PhysioNet Base-Model



(c) Own Base-Model

Fig. 7: Comparison of base-model training using the PhysioNet dataset (b) and own recordings (c).

depicts the average accuracy of a subject-independent model trained on PhysioNet tested on new subjects of our dataset. With a decline of 15% in accuracy from testing on the same dataset (66%) vs. testing on a different dataset (51%), it can be concluded that the learned knowledge from PhysioNet can not be transferred sufficiently to our small dataset with presumably too many recording differences.

For this reason, the more individualized fine-tuning approach is evaluated in the following section.

5.2.2. Fine-Tuning Evaluation

In order to understand which fine-tuning methods lead to the best classification results, the following combinations will be evaluated:

(a) PhysioNet base-model, fine-tuned on left out test-subjects of PhysioNet

(b) PhysioNet base-model, fine-tuned on target subjects of own dataset

(c) Own dataset base-model, fine-tuned on target subjects of own dataset

First, the subject-independent base model is evaluated on the unseen target subjects without fine-tuning (yellow), serving as a baseline. After the fine-tuning process, the adapted models are evaluated again with the validation set (blue). In

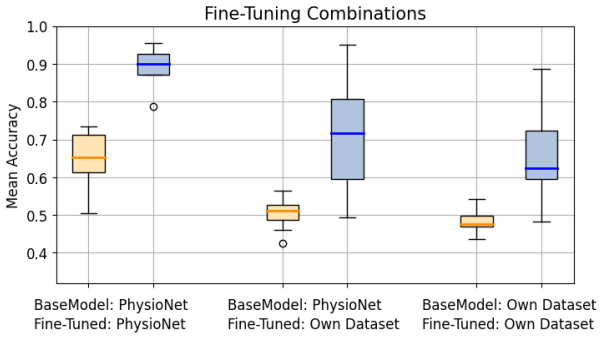


Fig. 8: Accuracies of different fine-tuning combinations, averaged over all target subjects. *Yellow*: Subject-independent evaluation of the base model without fine-tuning. *Blue*: Subject-specific evaluation of the adapted models after fine-tuning.

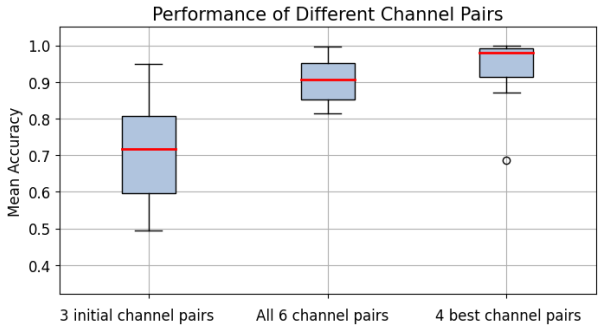


Fig. 9: Accuracies with different channel pair selections, averaged over fine-tuned models of all target subjects.

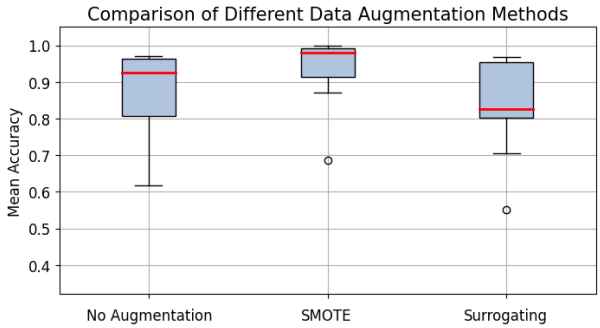


Fig. 10: Accuracies using different augmentation methods, averaged over fine-tuned models of all target subjects.

figure 8, the mean accuracies over all target subjects are displayed. Using the first combination, it can be seen that with a PhysioNet base model and fine-tuning on target subjects from the PhysioNet as well, high accuracies of 90% can be achieved. When fine-tuning on subjects from our dataset, the median accuracy is 71%. As expected, the performance is worse than with subjects of the same dataset. However, the improvement through fine-tuning is significant with an increase of 20% accuracy. When using a base model trained on our own data (right), worse results are achieved with a median accuracy of 62%. Therefore, it is recommended to train a subject-independent base model on a big dataset before fine-tuning it on target subjects of the small dataset.

Although a good accuracy can be achieved with the combination of the PhysioNet base model and own dataset fine-tuning, it is noticeable that the variance is very large, meaning that the accuracies of different target persons vary a lot. The source of this variance will be explored in the next section.

5.2.3. Channel Pair Selection

To gain more insights into the results of the fine-tuning training, a new PhysioNet base model as well as the individual fine-tuned models were trained with all 6 possible channel pairs, marked in different colors in figure 4. Table 3 states the mean accuracies of all individual channel pairs of each subject. It stands out, that for most subjects, samples that were recorded with the channel pairs [F3,F4] and [P3,P4] are classified with less accuracy. For subjects 1, 2, 10, and 11 the biggest differences are noticeable. A reason for that could be the bad placement and therefore bad connection between the electrodes and the skin. While recording the data, this could have caused more noise than with other electrode channels. As in the previous evaluation, the three channel pairs [C3,C4], [F3,F4], and [P3,P4] were used, the effect on the mean accuracy as well as the variation was significant. Compared to the previous channel pair selection, figure 9 gives insights into the fine-tuning performance on different channel settings.

Using all 6 channel pairs for fine-tuning, a median accuracy of 90% can be achieved with an improvement of almost 20% compared to the initial selection. Choosing the best four channel pairs [C3,C4], [Fp1,Fp2], [F7,F8] and [T3,T4], an even higher accuracy of 98% and a smaller variance is reached.

5.2.4. Evaluation of Augmentation Methods

Finally, the two different data augmentation methods SMOTE and surrogating are evaluated. Figure 10 gives insights into the average classification performance of the fine-tuned models using different augmentation on the training data. With a median accuracy of 92%, using no augmentation serves as a baseline. With any of the augmentation methods, the number of trials was multiplied by a factor of 4. Using SMOTE, a high median accuracy of 98% can be reached and a small variance is achieved. However, when surrogating the data, the median average of 82% is lower than the baseline. This leads to the conclusion that the SMOTE method is suggested, reaching the best results compared to no augmentation or surrogating.

Channel Pair	Test Subjects										
	Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8	Sub 9	Sub 10	Sub 11
$[C3, C4]$	0.96	0.91	0.99	0.98	1.0	0.97	1.0	1.0	0.98	0.82	0.87
$[F3, F4]$	0.95	0.77	0.74	0.63	0.46	0.99	1.0	0.99	1.0	0.91	0.96
$[P3, P4]$	0.67	0.70	0.91	0.86	0.79	0.53	0.90	1.0	0.92	0.60	0.49
$[Fp1, Fp2]$	0.97	0.91	0.99	0.95	0.95	0.99	1.0	1.0	1.0	0.74	0.92
$[F7, F8]$	0.97	0.92	1.0	0.97	0.97	0.99	1.0	0.99	0.99	0.91	0.95
$[T3, T4]$	0.93	0.88	1.0	0.97	0.93	0.97	1.0	1.0	0.91	0.91	0.80
Mean(All pairs)	0.91	0.85	0.94	0.89	0.85	0.91	0.98	1.0	0.97	0.82	0.83

Table 3: Detailed evaluation of fine-tuned models, separated into different channel pairs and subjects. Each accuracy is averaged over all test trials belonging to the respective channel-pair-subject combination.

5.2.5. Summary - Best Approach

After exploring different approaches, the approach that achieved the best classification accuracy of 98%, averaged over 11 target subjects of our own dataset is the following:

(1) Finding a big dataset with a similar/identical MI classification task (PhysioNet) and adapting it to our own recording settings (resampling, channel selection)

(2) Training a base model on the big dataset

(3) Collecting data with our own BCI system of a target subject

(4) Augmenting the collected data (SMOTE)

(5) Fine-Tuning the base-model with 80% of the subject-specific data and using the other 20% for evaluating the model

(6) Selecting channel pairs that lead to the most robust results, as depending on the BCI setting, some channels might not be advantageous in their positioning or are more prone to noise and connectivity issues

6. DISCUSSION AND FUTURE WORK

In this study, different approaches to classify samples of a small dataset of upper limb BCI-MI recordings are evaluated. After finding out, that classical EEG feature extraction methods such as CSP, combined with a classification of an SVM, reach their limitations when classifying data of multiple subjects, a 1D-CNN DL approach was explored. The focus was placed on how to exploit a large public existing data set to implement BCI projects with little existing data. The first key finding is, that the size of the dataset is crucial when training a subject-independent model with the goal to achieve high accuracies on unseen subjects. With an increase of 18% in accuracy, the model trained on the PyhsioNet dataset with 104 subjects performed significantly better than our model trained on 10 subjects. Based on the sufficient ability of the PhysioNet model to generalize on new data of the same dataset, the research focused on how this knowledge can be used for classifying our own data. The evaluation showed that insufficient results of 51% accuracy are achieved when training a base model on the big dataset PhysioNet and trying to transfer the knowledge directly to new data from different recording settings (our dataset). Retrieved from the too-large differences between the datasets, significantly better results can be achieved when adapting the model on the new dataset by fine-tuning it on our own subjects, reaching a median accuracy of 71%. However, after obtaining better results with the fine-tuned model, how much could the model benefit from the pre-trained PhysioNet model, or can the improved result only be attributed to the re-training on the own data? It was found, that despite the difficulties of transferring the knowledge of the PhysioNet base model to our data set, better results can be obtained than training and fine-tuning only with our own data. Thus, the pre-trained knowledge from the big dataset can be exploited by the fine-tuned model. Since the accuracy of the fine-tuned model has a noticeably large variance, the result was analyzed more in detail by evaluating the individual performance on samples of different channel pairs. It was found, that the four channel pairs [C3,C4], [Fp1,Fp2], [F7,F8] and [T3,T4] were classified most consistently with a high accuracy in all subjects. By changing the channel selection from the three initial pairs [C3,C4],[F3,F4] and [P3,P4] to the best four, the median accuracy could be improved from 71% to 98% with a smaller variance. Additionally, the two different augmentation methods SMOTE and surrogating were implemented, where SMOTE could lead to the best results.
In conclusion, training a base model on a big dataset, fine-tuning it with little data of one subject of our own dataset while making use of data augmentation, and selecting channel pairs that are most robust, can lead to a high classification accuracy of 98%.

The results of this work implicate that also small datasets can make use of DL methods to classify the subject's data reliably. One potential application of these findings is in the development of BCI systems for stroke patients who have lost motor function due to damage in the brain. The system would

allow them to control external devices, such as exoskeletons to perform the intended movement, using their thoughts. Subject-specific approaches, in which a large amount of data is collected per patient to train a specific model, are impractical. This would involve a lot of time and effort before the patient could use the BCI system for rehabilitation. In addition, the accuracy of these rehabilitation systems is critical, as even small errors can lead to frustration of the patient and ineffective control. The findings of this research suggest that before collecting data of a stroke patient, a base model can be trained on a big dataset, learning to extract features from EEG data. Making use of that pre-learned knowledge, only a small amount of data needs to be collected of the patient for fine-tuning a subject-specific adaption with minimal retraining time. The individualized BCI system could achieve a more accurate and reliable classification, providing a simple solution with only a few required EEG channels, that could help stroke patients regain some of their lost motor function, which can have a significant impact on their quality of life.

Comparing the findings of this work to previous research, similar improvement through fine-tuning can be stated. However, unlike previous research, where the fine-tuning subjects were taken from the same dataset as the base model was trained on, this work offers new insights into fine-tuning with a different dataset. These different starting conditions make the results of this work difficult to compare with other results from previous work. Nevertheless, the findings of this work provide valuable insights into the capabilities and limitations of classifying small BCI-MI datasets. As the results were promising, this approach should be further pursued and explored.

Major research potential is in expanding the dataset of 11 subjects to find more insights on the variety of data within a person. Since only data from one recording session per person was available in this dataset, the results are expected to be worse if the measurements were taken on different days. Data from different sessions bring more variety and thus more difficulty to the classification of the data due to the different applications of the BCI-cap, different states and concentration levels of the test subject, and minimal changes in the measurement environment. Therefore, handling more variable data is to be expected of a BCI system for rehabilitation. In addition, different model architectures could be evaluated. Especially the combination of CNNs and Transformer models has recently gained more attention for subject-independent EEG classification [19], [32]. Moreover, including a resting-state baseline class in training, as in the work of Mattioli et al., [33], could improve the rehabilitation system by not only differing between the left and right arm movements but also detecting when a movement should be performed in general.

7. REFERENCES

[1] Vladimir Hachinski, Geoffrey A Donnan, Philip B Gorelick, Werner Hacke, Steven C Cramer, Markku Kaste, Marc Fisher, Michael Brainin, Alastair M Buchan, Eng H Lo, et al., "Stroke: working toward a prioritized world agenda," Stroke, vol. 41, no. 6, pp. 1084–1099, 2010.

[2] Aekaterini Galimanis, Marie-Luise Mono, Marcel Arnold, Krassen Nedeltchev, and Heinrich P Mattle, "Lifestyle and stroke risk: a review," Current opinion in neurology, vol. 22, no. 1, pp. 60–68, 2009.

[3] Kai Keng Ang, Cuntai Guan, Karen Sui Geok Chua, Beng Ti Ang, Christopher Kuah, Chuanchu Wang, Kok Soon Phua, Zheng Yang Chin, and Haihong Zhang, "A clinical study of motor imagery-based brain-computer interface for upper limb robotic rehabilitation," in 2009 annual international conference of the IEEE engineering in medicine and biology society. IEEE, 2009, pp. 5981–5984.

[4] Muhammad Ahmed Khan, Rig Das, Helle K Iversen, and Sadasivan Puthusserypady, "Review on motor imagery based bci systems for upper limb post-stroke neurorehabilitation: From designing to application," Computers in biology and medicine, vol. 123, pp. 103843, 2020.

[5] Magdalena Ietswaart, Marie Johnston, H Chris Dijkerman, Sara Joice, Clare L Scott, Ronald S MacWalter, and Steven JC Hamilton, "Mental practice with motor imagery in stroke recovery: randomized controlled trial of efficacy," Brain, vol. 134, no. 5, pp. 1373–1386, 2011.

[6] GN Ranky and S Adamovich, "Analysis of a commercial eeg device for the control of a robot arm," in Proceedings of the 2010 IEEE 36th Annual Northeast Bioengineering Conference (NEBEC). IEEE, 2010, pp. 1–2.

[7] Francesco Carrino, Joel Dumoulin, Elena Mugellini, Omar Abou Khaled, and Rolf Ingold, "A self-paced bci system to control an electric wheelchair: Evaluation of a commercial, low-cost eeg device," in 2012 ISSNIP biosignals and biorobotics conference: biosignals and robotics for better and safer living (BRC). IEEE, 2012, pp. 1–6.

[8] Alexander Craik, Yongtian He, and Jose L Contreras-Vidal, "Deep learning for electroencephalogram (eeg) classification tasks: a review," Journal of neural engineering, vol. 16, no. 3, pp. 031001, 2019.

[9] Fabien Lotte, Camille Jeunet, Jelena Mladenović, Bernard N'Kaoua, and Léa Pillette, "A bci challenge for the signal processing community: considering the user in the loop," 2018.

[10] Qiyun Huang, Zhijun Zhang, Tianyou Yu, Shenghong He, and Yuanqing Li, "An eeg-/eog-based hybrid brain-computer interface: Application on controlling an integrated wheelchair robotic arm system," Frontiers in neuroscience, vol. 13, pp. 1243, 2019.

[11] Paula Sánchez López, Helle K Iversen, and Sadasivan Puthusserypady, "An efficient multi-class mi based bci scheme using statistical fusion techniques of classifiers," in TENCON 2019-2019 IEEE Region 10 Conference (TENCON). IEEE, 2019, pp. 378–382.

[12] Amjed S Al-Fahoum and Ausilah A Al-Fraihat, "Methods of eeg signal features extraction using linear analysis in frequency and time-frequency domains," International Scholarly Research Notices, vol. 2014, 2014.

[13] Daniel Planelles, Enrique Hortal, Álvaro Costa, Andrés Úbeda, Eduardo Iáñez, and José M Azorín, "Evaluating classifiers to detect arm movement intention from eeg signals," Sensors, vol. 14, no. 10, pp. 18172–18186, 2014.

[14] Saugat Bhattacharyya, Anwesha Khasnobish, Amit Konar, DN Tibarewala, and Atulya K Nagar, "Performance analysis of left/right hand movement classification from eeg signal by intelligent algorithms," in 2011 IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB). IEEE, 2011, pp. 1–8.

[15] Saugat Bhattacharyya, Anwesha Khasnobish, Somsirsa Chatterjee, Amit Konar, and DN Tibarewala, "Performance analysis of lda, qda and knn algorithms in left-right limb movement classification from eeg data," in 2010 International conference on systems in medicine and biology. IEEE, 2010, pp. 126–131.

[16] Navneet Tibrewal, Nikki Leeuwis, and Maryam Alimardani, "Classification of motor imagery eeg using deep learning increases performance in inefficient bci users," Plos one, vol. 17, no. 7, pp. e0268880, 2022.

[17] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization," Human brain mapping, vol. 38, no. 11, pp. 5391–5420, 2017.

[18] Hauke Dose, Jakob S Møller, Helle K Iversen, and Sadasivan Puthusserypady, "An end-to-end deep learning approach to mi-eeg signal classification for bcis," Expert Systems with Applications, vol. 114, pp. 532–542, 2018.

[19] Jin Xie, Jie Zhang, Jiayao Sun, Zheng Ma, Liuni Qin, Guanglin Li, Huihui Zhou, and Yang Zhan, "A transformer-based approach combining deep learning network and spatial-temporal information for raw eeg classification," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 30, pp. 2126–2136, 2022.

[20] Kaishuo Zhang, Neethu Robinson, Seong-Whan Lee, and Cuntai Guan, "Adaptive transfer learning for eeg motor imagery classification with deep convolutional neural network," Neural Networks, vol. 136, pp. 1–10, 2021.

[21] Hamdi Altaheri, Ghulam Muhammad, Mansour Alsulaiman, Syed Umar Amin, Ghadir Ali Altuwaijri, Wadood Abdul, Mohamed A Bencherif, and Mohammed Faisal, "Deep learning techniques for classification of electroencephalogram (eeg) motor imagery (mi) signals: A review," Neural Computing and Applications, pp. 1–42, 2021.

[22] Ping Wang, Aimin Jiang, Xiaofeng Liu, Jing Shang, and Li Zhang, "Lstm-based eeg classification in motor imagery tasks," IEEE transactions on neural systems and rehabilitation engineering, vol. 26, no. 11, pp. 2086–2095, 2018.

[23] Tian-jian Luo, Chang-le Zhou, and Fei Chao, "Exploring spatial-frequency-sequential relationships for motor imagery classification with recurrent neural network," BMC bioinformatics, vol. 19, no. 1, pp. 1–18, 2018.

[24] Syed Umar Amin, Mansour Alsulaiman, Ghulam Muhammad, Mohamed Amine Mekhtiche, and M Shamim Hossain, "Deep learning for eeg motor imagery classification based on multi-layer cnns feature fusion," Future Generation computer systems, vol. 101, pp. 542–554, 2019.

[25] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," Circulation, vol. 101, no. 23, pp. e215–e220, 2000 (June 13), Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

[26] Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J Miller, Gernot R Müller-Putz, Guido Nolte, Gert Pfurtscheller, Hubert Preissl, Gerwin Schalk, Alois Schlögl, Carmen Vidaurre, Stephan Waldert, and Benjamin Blankertz, "Review of the bci competition iv.," FRONT HUM NEUROSCI, vol. 6, pp. 55, 2012.

[27] Mark Wronkiewicz, Eric Larson, and Adrian KC Lee, "Leveraging anatomical information to improve transfer learning in brain–computer interfaces," Journal of neural engineering, vol. 12, no. 4, pp. 046027, 2015.

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[29] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al., "Language models are unsupervised multitask learners," OpenAI blog, vol. 1, no. 8, pp. 9, 2019.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[31] Jesse Vig, "A multiscale visualization of attention in the transformer model," arXiv preprint arXiv:1906.05714, 2019.

[32] Hamdi Altaheri, Ghulam Muhammad, and Mansour Alsulaiman, "Physics-informed attention temporal convolutional network for eeg-based motor imagery classification," IEEE Transactions on Industrial Informatics, vol. 19, no. 2, pp. 2249–2258, 2022.

[33] F Mattioli, C Porcaro, and G Baldassarre, "A 1d cnn for high accuracy classification and transfer learning in motor imagery eeg-based brain-computer interface," Journal of Neural Engineering, vol. 18, no. 6, pp. 066053, 2022.

[34] Rig Das, Paula S Lopez, Muhammad Ahmed Khan, Helle K Iversen, and Sadasivan Puthusserypady, "Fbcsp and adaptive boosting for multiclass motor imagery bci data classification: A machine learning approach," in 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2020, pp. 1275–1279.