2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Multi-subject classification of Motor Imagery EEG signals using transfer learning in neural networks

Carlos Emiliano Solórzano-Espíndola[1], Erik Zamora[1] and Humberto Sossa[1,2]

Abstract— **Brain-Computer Interfaces are new technologies with a fast development due to their possible usages, which still require overcoming some challenges to be readily usable. The paradigm of motor imagery is among the ones in these types of systems where the pipeline is tuned to work with only one person as it fails to classify the signals of a different person. Deep Learning methods have been gaining attention for tasks involving high-dimensional unstructured data, like EEG signals, but fail to generalize when trained on small datasets. In this work, to acquire a benchmark, we evaluate the performance of several classifiers while decoding signals from a new subject using a leave-one-out approach. Then we test the classifiers on the previous experiment and a method based on transfer learning in neural networks to classify the signals of multiple persons at a time. The resulting neural network classifier achieves a classification accuracy of 73% on the evaluation sessions of four subjects at a time and 74% on three at a time on the BCI competition IV 2a dataset.**

I. INTRODUCTION

The recent development in brain-computer interfaces (BCI) systems has allowed the identification of a range of techniques for the efficient decoding of several paradigms [1]. One of these is motor imagery, which refers to the systems aiming to decode the patterns in the brain activity exhibited when the subject is planning to execute a movement [2]. The practical applications of these technologies allow translating the activity into commands for systems that range from healthcare applications, like a wheelchair or prosthesis, to entertainment applications like video games [2].

While invasive methods have better resolution and less noise [3], electroencephalography (EEG) is one of the most common methods of acquisition of brain activity. EEG has the advantages of an appropriate time resolution, the electrode placement is flexible, and it does not require surgical intervention to place the device as invasive methods like electrocorticography (ECoG) [4].

A BCI pipeline is primarily composed of pre-processing to clean the signal from noise, followed by calculation of features to characterize the recording and reduce the dimension, and lastly a classification step to output the correct command. Considering EEG signals as the input, the signal filters are calculated for allowing the alpha (8-13 Hz) and beta (13-35 Hz) waves as the related patterns occur in this frequency bands [5].

Although using more data in machine learning tends to increase performance, the current approaches on motor imagery consider the subject-specific setup, thus limiting the training samples when more than one subject is available. Some explanations behind the inter-subject variability of sensorimotor rhythms (SMR) can be found in [6]. The motor learning process, brain function, and brain topology are some of the causes that can differ from person to person, leading to variability in the calculated features across different subjects. In the work of [7], the variability among subjects is explored with a pairwise inter-subject classification approach in the BCI competition IV 2a dataset training with one person and testing on another one. The best pair was for training on subject 3 and testing on subject 9 with 57.99%. In [8] a comparison of classifiers on the same dataset is done for the intra and inter-subject classification tasks finding that there is a low relationship ($R^2 = 0.153$) to the performance of the same classifier on both tasks.

An overview of techniques used on BCIs is described by F. Lotte et. al. [9], along with their advantages and applications. The work also establishes a division of techniques as classical machine learning approaches and new techniques such as adaptive classifiers, matrix classification, and deep neural networks.

In this paper, we first present feature extraction techniques of common spatial patterns (CSP) and filter-bank common spatial patterns (FBCSP), which have been used for the motor imagery task in the proposed dataset, and also a more recent approach based on Riemannian geometry. Then explore some of the machine learning classifiers, which are used as the last step to estimate the correct command, and propose an experiment to measure how every subject influences the classification to select candidates for multi-subject classification. Finally, a deep learning approach with the raw signals as the input is proposed to classify the signals of multiple subjects, where the first layers were pre-trained using autoencoders to reconstruct the signal.

II. METHODS

A. Data

The Berlin Brain-Computer Interface group has presented challenges with their respective datasets related to different

[1] Carlos Emiliano Solórzano-Espíndoa, Erik Zamora and Humberto Sossa, *Senior Member, IEEE*, are with Instituto Politécnico Nacional - CIC, Av. Juan de Dios Batiz S/N, Gustavo A. Madero, 07738 Mexico City, Mexico. csolorzanoe1200@alumno.ipn.mx, ezamorag@ipn.mx, hsossa@cic.ipn.mx

[2]Humberto Sossa is with Tecnológico de Monterrey, Campus Guadalajara. Av. Gral. Ramón Corona 2514 Zapopan, Jalisco. 45138, México.

H. Sossa and E. Zamora would like to acknowledge the support provided by CIC-IPN in carrying out this research. This work was economically supported by SIP-IPN (grant numbers 20200651, 20210316 and 20210788), CONACYT Fronteras de la Ciencia 65 and FORDECYT-PRONACES 60055. CE Solórzano-Espíndola acknowledges CONACYT for the scholarship granted towards pursuing his postgraduate studies.

classification setups on these systems. The BCI competition IV presented four challenges; a summary of these and the winner's solutions can be seen in [10]. Challenge 2 presents two datasets 2a and 2b of motor imagery EEG signals with ocular artifacts.

Dataset 2a is composed of 22 channels EEG recordings of four classes recorded on nine subjects [11]. The classes are left hand, right hand, tongue, and feet. Additionally, they include three channels of electrooculography (EOG) recordings to use noise reduction/suppression techniques when the noise source is known as independent component analysis (ICA). The recordings are split into two sessions recorded on different days; they are referred to as training and evaluation sessions. Each session has six runs, and each run has 48 trials for a total of 288 (144 for each class).

B. Pre-processing

The first step of the pre-processing is frequency-based filtering. In this case, the filters pass the frequency band from 8 to 35 Hz. The proposed filters have a total 16th order and a Butterworth response. As autoencoder architectures are proposed for the pre-training of the classifiers, the signals are scaled to fit within the range of the *tanh* activation function while ignoring extreme values. The scaling sets the 5th percentile value to -1 and the 95th percentile as 1.

C. Feature Extraction

The feature extraction outputs a feature vector $\phi(X) \in R^d$ to capture relevant information about the phenomena of interest. It also reduces the dimensionality of the data, as most of the machine learning approaches suffer from the curse of dimensionality [9].

CSP calculates linear combinations of the channels that maximize the variance of the signals of a class, and reduce it for the others [12], [13]. Filter Bank CSP (FBCSP) filters the EEG signal for a set of frequency ranges and applies CSP to each one of the resulting signals [2].

Riemannian geometry is useful for the matrix-based classification as mentioned on [9]. In this case, the tangent space approach maps the covariance matrices to a locally Euclidean space where they can be represented by a vector [14].

D. Classification

Classifiers are functions of the form $f : \phi(X) \to Y$ that map the features $\phi(X) \in \mathbb{R}^D$ to an estimated output $y \in Y$. Linear classifiers of the form $\hat{y} = f\left(\sum_i \beta_i \phi_i(X) + \beta_0\right)$ are the simpler models where each feature $\phi_i(X)$ is weighted by a corresponding β_i parameter to produce a hyperplane [9], [15]. Linear discriminant analysis (LDA), support vector machines (SVM) and logistic regressions (LR) are linear classifiers with different fitting methods. Multi-layer perceptrons (MLP) and random forest (RF) are non-linear classifiers that can find more complex decision boundaries through ensambles of functions [16], [17]. As the model complexity increases it is easier to learn more complex patterns but also overfit to the training data.

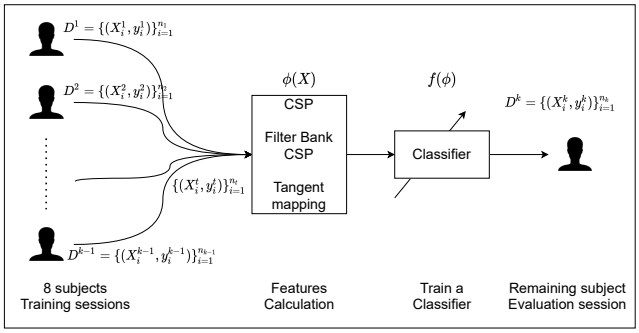


Fig. 1. Steps on the first experiment.

E. Deep Learning

Deep learning refers to neural network architectures with many layers that process the input as multidimensional tensors [9]. These layers process their inputs to benefit from the spatial, like the convolutional layers (ConvNets), or temporal, like the gated recurrent units (GRU), patterns of the data. These architectures have been found useful for tasks with unstructured data such as images or signals without the need for feature extraction functions [5].

Autoencoders are deep learning architectures that learn a compressed code through the encoder and learn how to reconstruct the original input from this code. As this code contains the information relevant to reconstruct the EEG signal, the encoder can be used as a pre-trained architecture that can be transferred to a classifier [5].

F. Experiments

1) Experiment 1: As stated before, the classification of multiple subjects depends on the inter-subject *associativity* among pairs [7]. In this work, we aim to limit the task to classify the signals of three and four subjects. Ideally, it is desirable to find the most compatible combinations for each case, as there are 84 combinations of three subjects and 126 of four. We limit this to use what we estimate to the subjects that can be best classified with classifiers fitted without signals of that subject.

The first experiment aims to measure the performance of a tuned pipeline for a set of subjects when tested on a new subject. For this, we tune a feature extraction function $\phi_k(X)$ and a classifier $f_k(\phi)$ using all the sample pairs (X_i^t, y_i^t) on the training sessions for each subject except for one k subject. The remaining subject's evaluation session is then used as the test set to measure how well their signals are classified using the information from others.

To find the best parameters 5-fold cross-validation is performed over a grid of the respective parameters for each classifier and the number of filters when they use CSP. Once that the best hyper-parameters are found, the classifier is trained using the whole training set and is used to report the training and test accuracy.

2) Experiment 2: In the second experiment, we use the best three and four subjects from the previous experiment to tune the feature extraction and classifiers functions and

TABLE I

ACCURACY ON THE TRAIN SET FOR THE PROPOSED MODELS FOR THE LEAVE-ONE-OUT TASK.

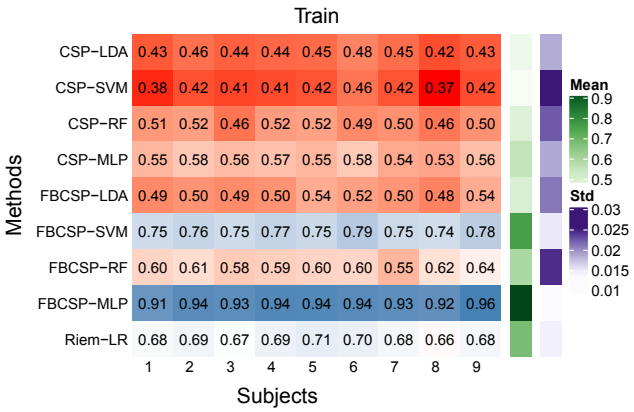
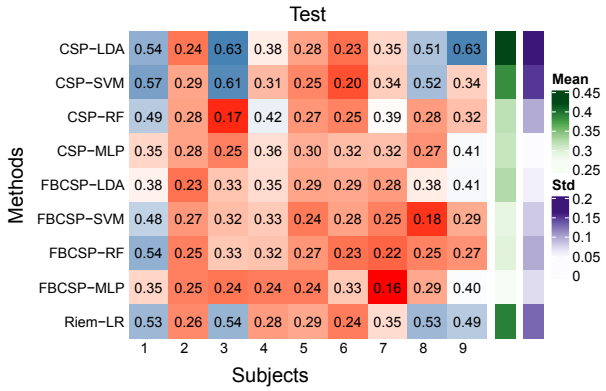


TABLE II

ACCURACY ON THE TEST SET FOR THE PROPOSED MODELS FOR THE LEAVE-ONE-OUT TASK.



test on these subjects again. In this case, we report three accuracies that are defined as follows:

- Train: The training sessions from the selected subjects.
- Test: The evaluation sessions from the selected subjects.
- Out: The remaining evaluation sessions.

The case of transfer learning differs as the autoencoder can use the training sessions from all nine subjects as a pre-training. To optimize the architecture for the autoencoder, a set of candidate filter dimensions and the number of filters per layer are proposed. To evaluate the architecture the reconstruction error is measured as the mean squared error (MSE) on the evaluation sessions. The final classifier uses a combination of the number of layers transferred from the autoencoder, the number of GRU cells, and the remaining fully connected (FC) layers. The best architecture is validated using a subset with 20% of the training set.

We also train the architecture with and without data augmentation (DA) as described in [18] to randomly select and crop a signal in the training set to have more samples. Additionally, noise from the EOG signals is randomly cropped and added to the samples.

III. RESULTS

A. Leave-one-out

The results from the first experiment are on the tables I and II for the training and test accuracy with all the subject-model pairs. The overall mean of the training accuracy is 0.6 and 0.34 for the test accuracy. There are some instances where the test accuracy is higher than the training accuracy, especially in the cases with the highest test accuracy. This is expected as the classifiers are trained to classify the signals from multiple subjects but are tested only on one. The classifiers with the best test accuracy (CSP-LDA, CSP-SVM, Riem-LR) are the ones using linear classifiers.

The distributions of the accuracy of the models categorized per each subject can be seen in Fig. 2, and are summarized in the Table III. The reported p-value is calculated from a one-side student's t-test comparing the resulting accuracies

TABLE III

STATISTICS OF THE TEST ACCURACY FOR EACH SUBJECT ACROSS THE CLASSIFIERS.

Subject	1	2	3	4	5	6	7	8	9
Average	**0.47**	0.26	0.38	0.33	0.27	0.26	0.30	0.36	0.40
STD	0.09	0.02	0.17	0.05	0.02	0.04	0.07	0.13	0.11
p-value	**3.7e-5**	0.08	0.02	7.7e-4	9.2e-3	0.17	0.058	0.021	2.2e-3

from that person to that of a random classifier ($\mu_0 = 0.25$) considering the alternative hypothesis of a higher mean. The best-found subjects are 1,9,3 and 8 according to their resulting averages across the models and having a p-value under 0.05.

B. Multi-subject classification

The results from this task are in Table V. The classifier with the best accuracy is the transfer model architecture with a test accuracy of 0.74 on the three subjects task and 0.73 on the four subjects. Using data augmentation on the architecture did not improve the performance and resulted in lower training and testing accuracy in both cases. Using this methodology improves the classification compared to the previous experiment with averages of 0.63 and 0.61 across all the proposed classifiers. The results on the out accuracy are

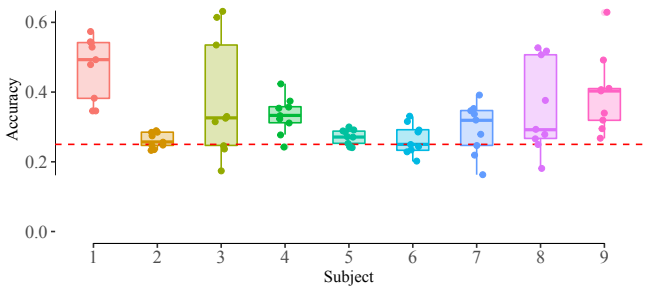


Fig. 2. Distribution of the reported accuracies for each subject on experiment 1. The dashed line is used as reference of a random classifier.

TABLE IV

ARCHITECTURE USED FOR THE MULTI-SUBJECT CLASSIFICATION TASK.

Layer	Units	Kernel Size	Activation	Regularization
Conv 1D	32	16	ELU[a]	BN(0.1)[b]
Conv 1D	64	12	ELU	BN(0.1)
Conv 1D	128	12	ELU	BN(0.1)
GRU	32		Linear	Dropout(0.2)
FC	32		ELU	Dropout(0.5)
FC	4		SoftMax	

[a]ELU: exponential linear unit
[b]BN: batch normalization

TABLE V

ACCURACY ON MULTI-SUBJECT CLASSIFICATION TASK.

	1,3,9			1,3,8,9		
	Train	Test	Out	Train	Test	Out
CSP-LDA	0.45	0.48	0.27	0.52	0.58	0.25
CSP-SVM	0.73	0.69	0.25	0.74	0.61	0.25
CSP-RF	0.64	0.52	0.25	0.62	0.51	0.25
CSP-MLP	0.66	0.59	0.25	0.73	0.64	0.25
FBCSP-LDA	0.72	0.65	0.25	0.68	0.57	0.25
FBCSP-SVM	0.91	0.71	0.25	0.90	0.68	0.25
FBCSP-RF	0.68	0.59	0.25	0.65	0.56	0.25
FBCSP-MLP	0.83	0.67	0.25	0.84	0.62	0.25
Riem+LR	0.75	0.71	0.25	0.76	0.69	0.25
Transfer Model	1.00	**0.74**	**0.34**	1.00	**0.73**	**0.29**
Transfer +DA	0.74	0.64	0.31	0.98	0.64	0.27
Average	0.74	0.63	0.26	0.74	0.61	0.25

those of a random classifier, except for the transfer models. Those results may be explained as the training sessions from all nine subjects were used for the autoencoder pre-training.

The final autoencoder has four layers of 1D convolutional layers that filter through time. This architecture achieves an MSE of 5.35 on the training sessions and 5.94 on the test sessions. For the classifier, the best architecture took three pre-trained layers from the encoder, followed by a recurrent layer(GRU) and two more fully connected layers. The training is stopped once that the validation accuracy increases instead of decreasing. The architecture is detailed in Table IV.

IV. CONCLUSION

In this work, we first introduced a benchmark of the classifier's performance to classify the signals of a specific subject with parameters calculated with the rest of the subjects on the dataset. The results show that the performance of the classifiers depends upon the person. It has also demonstrated that the features can classify the signals from subjects with better performance than a random classifier but still with low accuracy. From this test, we extracted potential candidates to test the multi-subject classification task.

In the multi-subject classification, using data from all the subjects to train an autoencoder and then transferring the layers to a classifier helps to leverage the problem of training a deep network with a small dataset using the raw signal as the input. As the final tuning is performed with what we found to be a limited set of candidate subjects, it achieves accuracy similar to classifiers tuned for a single subject on the four-class classification task and outperforms methods that require feature extraction.

REFERENCES

[1] S. N. Abdulkader, A. Atia, and M.-s. M. Mostafa, "Brain computer interfacing: Applications and challenges," *Egyptian Informatics Journal*, vol. 16, no. 2, pp. 213–230, jul 2015.

[2] S. Aggarwal and N. Chugh, "Signal processing techniques for motor imagery brain computer interface: A review," *Array*, vol. 1-2, no. June, p. 100003, 2019.

[3] K. J. Miller, D. Hermes, and N. P. Staff, "The current state of electrocorticography-based brain-computer interfaces," *Neurosurgical Focus*, vol. 49, no. 1, pp. 1–8, 2020.

[4] J. Minguillon, M. A. Lopez-Gordo, and F. Pelayo, "Trends in EEG-BCI for daily-life: Requirements for artifact removal," *Biomedical Signal Processing and Control*, vol. 31, pp. 407–418, 2017.

[5] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: a systematic review," *Journal of Neural Engineering*, vol. 16, no. 5, p. 051001, aug 2019.

[6] S. Saha and M. Baumert, "Intra- and Inter-subject Variability in EEG-Based Sensorimotor Brain Computer Interface: A Review," *Frontiers in Computational Neuroscience*, vol. 13, no. January, pp. 1–8, 2020.

[7] S. Saha, K. I. U. Ahmed, R. Mostafa, L. Hadjileontiadis, and A. Khandoker, "Evidence of variabilities in eeg dynamics during motor imagery-based multiclass brain–computer interface," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 2, pp. 371–382, 2018.

[8] C. E. Solórzano-Espíndola, H. Sossa, and E. Zamora, "A comparison study of eeg signals classifiers for inter-subject generalization," in *Pattern Recognition*, E. Roman-Rangel, Á. F. Kuri-Morales, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, and J. A. Olvera-López, Eds. Cham: Springer International Publishing, 2021, pp. 305–315.

[9] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for EEG-based brain-computer interfaces: A 10 year update," *Journal of Neural Engineering*, vol. 15, no. 3, 2018.

[10] M. Tangermann, K. R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. J. Miller, G. R. Müller-Putz, G. Nolte, G. Pfurtscheller, H. Preissl, G. Schalk, A. Schlögl, C. Vidaurre, S. Waldert, and B. Blankertz, "Review of the BCI competition IV," *Frontiers in Neuroscience*, vol. 6, no. JULY, pp. 1–31, 2012.

[11] C. Brunner, R. Leeb, G. R. Müller-Putz, A. Schlögl, and G. Pfurtscheller, *BCI Competition 2008 – Graz data set A*, 1st ed., Institute for Knowledge Discovery, Graz University of Technology, Graz, Austria, 1 2008.

[12] C. D. Virgilio Gonzalez, J. H. Sossa Azuela, E. Rubio Espino, and V. H. Ponce Ponce, *Classification of Motor Imagery EEG Signals with CSP Filtering Through Neural Networks Models*. Springer International Publishing, 2018, p. 123–135.

[13] C. Park, C. C. Took, and D. P. Mandic, "Augmented complex common spatial patterns for classification of noncircular eeg from motor imagery tasks," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 1, pp. 1–10, 2014.

[14] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Multiclass brain–computer interface classification by riemannian geometry," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 4, pp. 920–928, 2012.

[15] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY: Springer New York, 2009, no. 4.

[16] F. Arce, E. Zamora, G. Hernández, J. M. Antelis, and H. Sossa, "Recognizing motor imagery tasks using deep multi-layer perceptrons," in *Machine Learning and Data Mining in Pattern Recognition*, P. Perner, Ed. Cham: Springer International Publishing, 2018, pp. 468–482.

[17] D. Steyrl, R. Scherer, J. Faller, and G. R. Müller-Putz, "Random forests in non-invasive sensorimotor rhythm brain-computer interfaces: A practical and convenient non-linear classifier," *Biomedizinische Technik*, vol. 61, no. 1, pp. 77–86, 2016.

[18] Z. Tayeb, J. Fedjaev, N. Ghaboosi, C. Richter, L. Everding, X. Qu, Y. Wu, G. Cheng, and J. Conradt, "Validating deep neural networks for online decoding of motor imagery movements from eeg signals," *Sensors (Switzerland)*, vol. 19, no. 1, 2019.