ERP-based Brain-Computer Interface classification using different Machine Learning models

Aigerim Keutayeva
School of Engineering and Digital Science (SEDS)
Nazarbayev University (NU)
Astana, Kazakhstan
aigerim.keutayeva@nu.edu.kz

Rasul Yermagambet
School of Engineering and Digital Science (SEDS)
Nazarbayev University (NU)
Astana, Kazakhstan
rasul.yermagambet@nu.edu.kz

***Abstract*—Brain-computer interface (BCI) enables to communicate people with computers by processing electroencephalography (EEG) signals. One of the most popular paradigm is event related potential (ERP), that stimulates brain activity by using external event stimuli. Different people (subjects) have various responses of P300 ERP component, which is the major feature for target signal detection. Two ensemble learning classifiers based on weighted average method are proposed in order to produce robust results for many subjects utilizing generic training (GT). The models utilized to create ensemble algorithms are linear discriminant analysis (LDA), linear regression (LR), support vector machine (SVM), k-nearest neighbor (kNN) models. The proposed models are trained and tested on four datasets including ALS, NU, BNCI, EPFLP. The model based on LDA-SVC-kNN outperformed other models on two datasets with highest f-score of 94.31.**

***Index Terms*—Brain-computer interface, ERP, Moabb datasets, Machine Learning**

I. Introduction

People can communicate with computers or robots via a brain-computer interface (BCI), a system that uses electrical, magnetic, or hemodynamic neural impulses to infer users' cognitive states. Numerous BCI system designs have been built on the experimental paradigm that Farwell and Donchin described. In this paradigm, visual stimuli are presented to the user to generate event-related potentials (ERPs), signals obtained through electroencephalography (EEG). The ERP patterns, especially P300 waves, are subsequently converted to the proper control commands by the BCI system. The post-synaptic activity of pyramidal neurons in the brain is combined in the scalp-recorded ERPs. In other words, they record a spatial or temporal combination of background neural activity or visual/auditory/haptic sensory stimulus correlating to diverse background neuronal activities.

As a result, acquired ERPs are obscured by brain activity unrelated to the task and typically have a low signal-to-noise ratio. Furthermore, ERP waveforms show much variation. Data distribution fluctuates between sessions and even from trial to trial for the same participant. This situation is typically related to either the subject's neurophysiological state (such as the amount of exhaustion, stress, and sleepiness) or specific experimental conditions (e.g., electrode positions in sessions, head sweating, or loud surroundings). The challenges brought on by this kind of variability are made worse by the wide range of waveforms recorded from various subjects (i.e., intersubject variability).

EEG data collection, preprocessing, feature extraction, and classifier design based on the retrieved features are some critical machine-learning phases that convert brain events into commands for BCI devices. There are several studies on each stage, notably those that discuss deep learning classifiers for machine learning in BCI. Finding precise deep-learning models for ERP-based BCIs is still a challenge despite the efforts of recent studies. This study compares the outcomes of ensemble learning to several built-in models on preprocessed datasets.

The paper is organized as follows. In Section II, we describe publicly available datasets used in our study. Section III presents different machine learning models used. Section IV discusses the results. Section VI summarizes our key observations, and finally, Section VII concludes the paper.

II. Datasets

We used the Nazarbayev University dataset with other publicly available datasets from three different sources, as described below. These datasets are easily accessible via the MOABB tool [4].

A. NU Dataset

The Nazarbayev University (NU) dataset [1] is collected from 10 healthy subjects (five women and five men, aged 22-35 years) without any neurological, psychological, or physical disorders performing a visual P300 task for spelling. All participants had no prior experience related to BCI experiments. The dataset represents a record of P300 evoked potentials using the Farwell and Donchin paradigm. The EEG data were acquired using a g.USBamp, g.LADYbird, g.Tec, Austria at a sampling frequency of 256 Hz and a 16-channel with active Ag/AgCl electrodes. The ground electrode was placed on the participants' right earlobe, whereas a reference electrode location was set to FCz Fig. 1 shows the sensor locations used for the study.

B. ALS Dataset

The amyotrophic lateral sclerosis (ALS) dataset is collected from 8 subjects with ALS (three women and five men, age 58

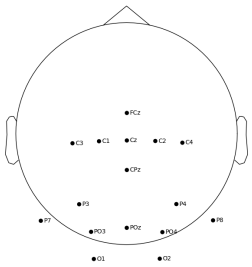

Fig. 1. EEG sensor locations and labels for NU dataset (16 channels)

± 12) performing a visual P300 task for spelling. The dataset represents a record of P300 evoked potentials using the Farwell and Donchin paradigm. The EEG data were acquired using a g.MOBILAB (g.tec, Austria) at a sampling frequency of 256 Hz and eight sensor electrodes were referenced to the right earlobe. Fig. 2 shows the sensor locations used for the study. The results are published in a paper by A Riccio et al. [6].

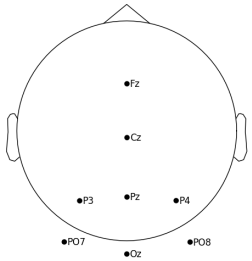

Fig. 2. EEG sensor locations and labels for ALS and BNCI datasets (8 channels)

C. EPFLP Dataset

The EEG data in this dataset (EPFLP300) was recorded at a 2048 Hz sampling rate from 32 electrodes placed at the standard positions of the 10-20 international system (see Fig. 3). The system was tested on four disabled and four healthy subjects. The disabled subjects were all wheelchair-bound but had varying communication and limb muscle control abilities (subjects 1 to 4). Subjects 5 to 8 were Ph.D. students recruited from the laboratory (all male, age 30 ± 2.3). The stimuli were six different images (a lamp, a television, a door, a window, a telephone, and a radio) flashed randomly with a 400-ms stimulus interval. Each subject completed four recording sessions in 2 days within two weeks. One session comprised, on average, 810 trials, and the whole data for one subject consisted of 3240 trials. The results are published in a paper by Hoffmann et al. [3].

D. BNCI Dataset

This dataset was collected using two different experimental paradigms based on overt and covert attention conditions on P300 Speller. This dataset (BNCI2015003) contains recordings from 10 subjects (age: 27.9 ± 10.9) performing a visual

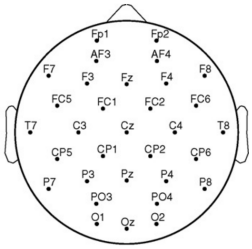

Fig. 3. EEG sensor locations and labels for EPFLP dataset (32 channels)

P300 task for spelling. Data were acquired by g.tec Medical Engineering GmbH, Austria, and Fondazione Don Gnocchi, Italy. The subjects were free of medication and central nervous system abnormalities and had no prior experience with EEG-based communication systems. The EEG data were acquired using a g.USBamp at a sampling frequency of 256 Hz and 8 electrodes were referenced to the right earlobe. Fig. 2 shows the eight electrode locations used for the study. High-pass and low-pass filters were used with cutoff frequencies of 0.1 and 20 Hz, respectively. Each subject participated in only one session. Results were published in a paper by C. Guger et al. [2].

III. Exploratory Data Analysis (EDA)

All four datasets used in this paper are different in shape, and in other aspects. In order to choose a model to train, the details must be explored. However, in this section we will discuss only ALS and BNCI datasets, as they are totally different because the patients are of different health conditions.

A. ALS Data

This dataset consists data signals from 8 subjects with ALS, where signals are as in the figure 4.

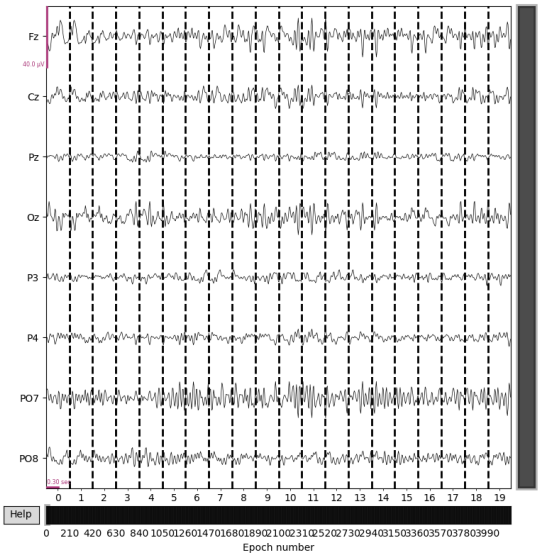

Fig. 4. ALS original data signals

Surprisingly the figure 4 shows less noise than in any other datasets. There are many reasons for that, including outside

noise, muscle control (which is impossible for ALS patients), visual noise, power noise, and etc.

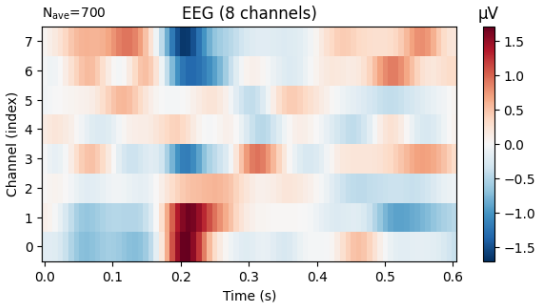

Fig. 5. ALS data signal amplitude in all 8 channels for subject1.

The figure 5 also shows much more better results than other datasets, but at the same time it has high channel-to-channel variability, or noise.

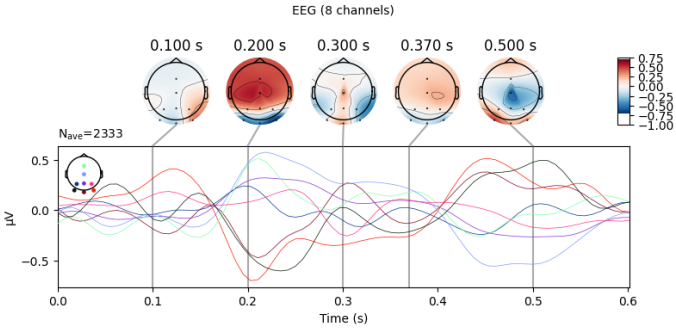

Fig. 6. ALS evoked signals for combined target and non-target signal of subject1

We can also take a look on evoked signals, that can be useful for noise detection too. The figure 6 ahows that the max aplitude is around 200ms, and the are a lot of noise in other temporal features.

This noise can be reduced by ICA Decomposition application 7.

B. BNCI Data

This dataset consists data from 10 subjects, where the trial number is 5400 for the first two subjects, while for others 1800. The data used only 8 channels within 128 ms temporary features. All ten subjects can be considered as healthy, that is why the results might be better than the previous datasets.

In the figure 8 we can see the difference of signals over different trials and average line. From the figure it is possible to note that trial-to-trial variability is high, which might decrease classification accuracy not only for pooled data, but also for subject specific cases.

From the figure 9 we can see another problem, the channel-to-channel variability is also high enough to affect on the accuracy. The data is unbalanced, it might be better to use f1 score, instead of accuracy metric.

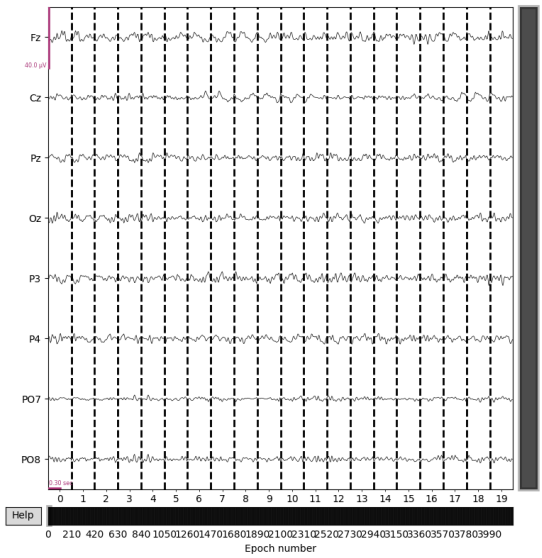

Fig. 7. ALS data signals after ICA.

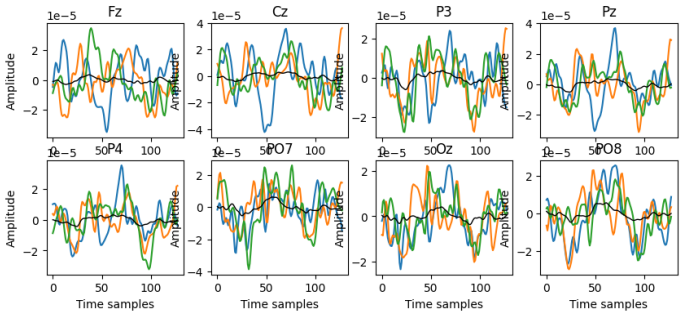

Fig. 8. BNCI subject1 data signal in three trials and average over all trials (black)

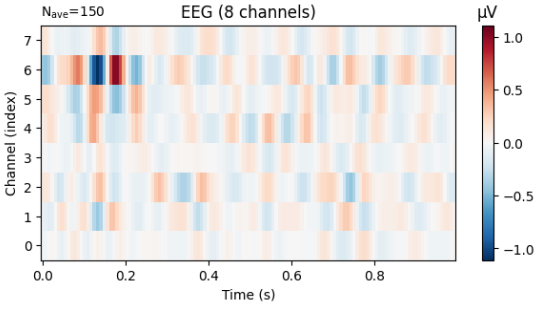

Fig. 9. BNCI subject1 data signal amplitude in all 8 channels

In order to solve one of the problems at least, we can apply ICA Decomposition and see the results in the figures 10 and 11.

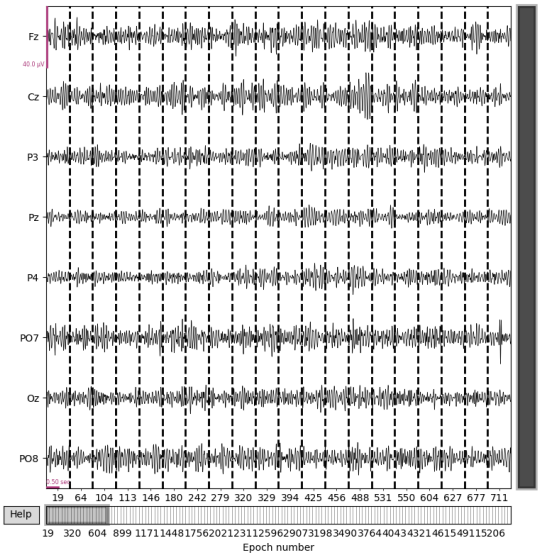

Fig. 10. BNCI original data signals

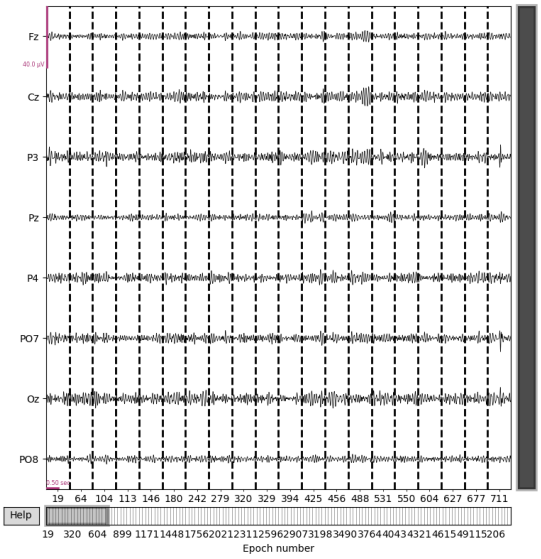

Fig. 11. BNCI data signals after ICA

We can absolutely see the difference, that ICA Decomposition helped with some noise existed.

IV. Model selection and Training

Table I shows the total number of observations in training, validation, and test sets across both target and non-target classes for each of the four datasets used in our model selection process.

In this paper, each dataset was trained in two ways: first, models were trained on a pooled sample (general) of subjects within each dataset; second, models were trained for each subject specifically within a dataset. The reason for that, is a difference of signals from trial to trial, from subject to subject, even from session to session. Therefore it was important to test information from different sides.

TABLE I
THE TRAINING, VALIDATION, AND TEST SETS SHOWN AS 3D ARRAYS (NUMBER OF TRIALS X NUMBER OF CHANNELS X TIME SAMPLES)

Datasets	Training Set	Validation Set	Test Set
NU data	(45333 x 16 x 76)	(5567 x 16 x 76)	(5567 x 16 x 76)
ALS data	(26880 x 8 x 78)	(3360 x 8 x 78)	(3360 x 8 x 78)
EPFLP data	(21312 x 32 x 77)	(2664 x 32 x 77)	(2664 x 32 x 77)
BNCI data	(20160 x 8 x 128)	(2520 x 8 x 128)	(2520 x 8 x 128)

TABLE II
THE TRAINING AND VALIDATION ACCURACY, F MEASURE (F1 SCORE) AND VARIANCE SCORES FROM VALIDATION DATA.

Datasets	Models	Train Acc	Valid Acc	Valid F1	Variance
NU data	KNN	100	85.94	3.86	-0.007
	LR	85.79	85.71	7.11	-0.042
	LDA	85.83	85.37	9.00	-0.085
	RF	100	85.95	1.97	0.001
	SVC	86.68	85.88	0	0
	MLP	98.21	79.54	27.15	-0.686
ALS data	KNN	100	83.42	90.86	-0.048
	LR	83.33	83.45	90.97	0
	LDA	85.28	84.82	91.42	-0.025
	RF	100	83.96	91.18	0
	SVC	86.25	84.32	91.36	0.020
	MLP	83.33	83.45	90.98	0
EPFLP data	KNN	100	83.71	90.99	-0.063
	LR	65.83	63.59	75.75	-1.454
	LDA	84.54	84.08	91.11	-0.069
	RF	99.99	83.82	91.15	-0.017
	SVC	83.36	83.74	91.15	0
	MLP	80.99	81.23	89.54	-0.273
BNCI data	KNN	100	88.33	93.80	-0.089
	LR	89.47	89.04	94.20	0.001
	LDA	89.70	87.70	93.42	-0.177
	RF	100	89.00	94.18	0
	SVC	90.53	89.00	94.18	0
	MLP	90.54	88.41	93.84	-0.020

A. Generic Training

Generic training approach combines the data from all separate subjects into one set. This approach allows to use the trained model on a new subject without additional training. The built-in scikit-learn classifiers, such as k-nearest neighbors algorithm (KNN), Logistic Regression (LR), Linear Discriminant Analysis (LDA), Random Forest (RF) Classifier, Support Vector Classifier (SVC), and Multi-layer Perceptron classifier were chosen for training. For ensemble learning, we used only LDA, SVC, kNN and LR. The metrics used are accuracy, f measure, and variance, as f measure and variance are suitable for unbalanced data as in our datasets. The results are in the following table II.

The results show accuracy between 80 and 90, and f measure between 85 and 95. It is advisable to change a model or apply some methods to increase these results. Therefore, we first applied cross-validation, and the results are shown in Table III.

TABLE III
THE CROSS-VALIDATION SCORES FOR CV=5, STRATIFIED CV, AND SHUFFLED CV

Datasets	Models	Simple CV=5	Stratified	Shuffled
NU data	KNN	85.31	85.3	85.39
	RF	85.40	85.39	85.39
	SVC	85.53	85.53	85.57
ALS data	LDA	83.87	83.87	84.71
	RF	83.51	83.51	83.61
	SVC	83.85	83.85	83.86
EPFLP data	LDA	81.14	81.14	82.94
	RF	82.75	82.63	83.32
	SVC	83.33	83.33	82.94
BNCI data	LR	87.33	87.33	89.29
	RF	77.77	77.72	89.21
	SVC	83.33	83.33	8910

Overall, cross-validation did not change the performance much, and somewhere it is even decreased. Moreover, model training with cross-validation took around 13-15 hours for each dataset with six models (more than 50 hours in total). However, we decreased the number of models to 3 and checked again. This time it took 5-6 hours each (more than 20 hours in total).

B. Ensemble Learning

Ensemble learning is a method that use combination of predictions from multiple algorithms in order to obtain better predictive performance. This method generally results in better performance, but requires more computational time. The type of ensemble learning used in this paper is averaged voting.

$$P_{avg}(X|y=1) = \frac{\sum_{i=1}^{N} P_i(X|y=1)}{N} \quad (1)$$

where $P_i(X|y=1)$ is the i th classifier's prediction of EEG vector X containing target P300 component. N is the ensemble voter's classifier count.

According to [5] the weighted ensemble learning performs better than simple ensemble averaging, but some cases show that this is not always true for CNN, SVM and LDA. Mussabayeva [5] uses weighted voting LDA-SVM-kNN classifier to achieve the best performance, where weights are picked using random search algorithm. It is claimed that the RS algorithm takes 40.89s to find the optimal weights, while training of the ensemble model itself took only 3.81s. This number significantly differs in our experiment. Time elapsed for training of only SVM on ALS dataset is 1805.97s for all subjects. However, we didn't perform random search to find the optimal weight parameters due to two reasons. Firstly, our initial goal was to propose another method for weighted average voting. Secondly, as the perfomance of the model was assessed on four datasets, we were limited by time and computational power. Our approach is aimed to use the accuracy scores from separately trained models as the weights for the ensemble voting. For instance, in Table V the accuracy test scores for LDA, SVC and kNN are 0.8478, 0.84122 and 0.7957 respectively. So basically if the model performance better in separate training, it gets more weight in voted ensemble model. In addition with accuracy, elapsed time, f-measure and variance scores are used to assess the models performance. Elapsed time is simply the time taken to train the model on the particular dataset. F-score or f-measure is also used to evaluate the accuracy of the model and it combines precision and recall values in it. Therefore, it is considered to be as an efficient metrics to use on unbalanced data. The variance score shows the dispersion of errors of a particular dataset. Values close to 1 are considered to be the good variance score.

$$F - measure = \frac{2*(Precision*Recall)}{Precission+recall} \quad (2)$$

Firstly, the same combination of ML models with the same hyperparameters as in [5] was implemented to build ensemble model. The best test accuracy was performed by LDA for ALS dataset which is 84.776%. Performance of SVC is 84.122 and kNN is 79.568. The ensemble model obtained the test accuracy of 84.776 comparing to [5] 85.35%, which is basically just the best result of LDA model. As it can be easily noticed, LDA performed the best despite the linearity of its nature and took only 2.37 seconds to train, while kNN performed the worst and SVM took 1805 seconds to be trained in comparison. W-LDA-SVC-kNN achieved the highest f-score and the best variance among the other models for ALS dataset. Our model performed better in value of f-score comparing to [5] As the linear models achieved the best result it was notioned to try another set of linear models to increase accuracy of the ensemble and decrease training time. We substituted kNN with Linear Regression and set SVM kernel to linear. Linear Regression performed better than kNN, but the new weights haven't increased the performance of the model. However, the time elapsed for the training of an ensemble model decreased three times, while accuracy remained the same. Moreover, f-score and variance of the new ensemble model slightly decreased.

The results of other 3 datasets interestingly varied from our expectations. It should be mentioned first, that kNN takes the least amount of time to train for all datasets. LDA achieved the highest score in terms of accuracy, f-score and variance for NU data. Performance of W-LDA-LSVC-LR model was better than W-LDA-SVC-kNN in all metrics types for the same dataset. It can be noted that this time difference between elapsed time for ensemble models is much smaller (125 s). For BNCI data, the best performance was achieved by linear regression in terms of accuracy and f-score. W-LDA-SVC-kNN showed the leading result in terms of variance. It is easy to notice, that this time the time elapsed to train both ensemble models differs only by 11 seconds. Finally, LDA slightly outperformed other models in terms of accuracy, while the highest f-score was achieved by W-LDA-SVC-kNN for EPFLP data. However, it took less time to train W-LDA-SVC-kNN than W-LDA-SVC-kNN. Linear Regression demonstrates dreadful results due to limitations of number of iterations for this dataset.

TABLE IV
THE SPECIFIC SUBJECTS VALIDATION SCORES FOR SPECIFIC MODELS

Datasets	Models	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	Average
NU data	KNN	85.70	85.84	85.64	86.11	86.32	85.41	84.97	84.68	85.14	85.80	85.56
	RF	85.45	85.59	85.64	86.11	86.91	86.16	88.41	85.10	85.63	85.68	86.07
	SVC	85.70	85.97	85.64	87.73	91.18	86.54	90.99	85.53	85.75	85.68	87.07
ALS data	LDA	85.36	85.12	85.95	85.71	87.02	87.50	87.14	91.67	-	-	86.93
	RF	83.10	85.60	83.21	83.21	83.81	83.81	85.60	85.00	-	-	84.17
	SVC	83.33	85.60	83.93	83.57	84.29	85.60	85.83	89.52	-	-	85.21
EPFLP data	LDA	79.97	81.80	86.25	84.94	82.42	86.07	90.42	83.08	-	-	84.37
	RF	86.25	83.16	87.44	83.13	85.91	84.74	86.38	82.78	-	-	84.97
	SVC	83.41	83.30	83.41	83.28	83.33	83.41	83.23	83.38	-	-	83.34
BNCI data	LR	95.46	96.76	80.83	83.61	82.22	78.89	81.39	86.94	85.28	87.50	85.89
	RF	97.22	97.22	84.17	83.33	83.61	83.06	82.78	83.33	83.06	83.33	86.11
	SVC	97.22	97.22	83.89	84.17	84.17	83.33	83.33	83.33	83.33	84.17	86.42

TABLE V
PERFORMANCE OF ENSEMBLE MODELS

Datasets	Models	Valid Acc	Valid F1	Variance	Elapsed Time
	LDA	84.78	91.39	-0.021	2.37
	SVC	84.12	91.26	0.019	1805.96
	kNN	79.57	88.39	-0.409	0.013
ALS data	Linear SVC	83.33	90.91	0.0	630.74
	LR	83.33	90.91	0.0	2.88
	W-LDA-SVC-kNN	84.78	91.52	0.024	1828.56
	W-LDA-LSVC-LR	84.78	90.91	0.0	634.40
	LDA	89.03	93.55	0.218	3.06
	SVC	88.05	93.21	0.202	400.98
	kNN	83.19	90.44	-0.151	0.012
NU data	Linear SVC	83.33	90.91	0.0	303.45
	LR	83.33	90.91	0.0	3.03
	W-LDA-SVC-kNN	88.92	90.91	0.0	425.47
	W-LDA-LSVC-LR	88.92	92.14	0.115	305.53
	LDA	88.31	93.76	-0.144	3.33
	SVC	89.29	94.34	0.0	523.68
	kNN	83.67	91.01	-0.698	0.016
BNCI data	Linear SVC	89.29	94.34	0.0	535.52
	LR	89.33	94.36	0.0	4.66
	W-LDA-SVC-kNN	89.25	92.14	0.115	549.05
	W-LDA-LSVC-LR	89.25	94.31	-0.013	538.30
	LDA	83.48	90.75	-0.087	10.56
	SVC	83.33	90.91	0.0	2117.06
	kNN	81.06	89.28	-0.288	0.037
EPFLP data	Linear SVC	83.33	90.91	0.0	2151.8
	LR	58.11	71.09	-1.675	13.32
	W-LDA-SVC-kNN	83.42	94.31	-0.013	2186.67
	W-LDA-LSVC-LR	83.26	90.83	-0.020	2202.30

C. Subject specific

Simple machine learning, including kNN, LR, LDA, RF, and SVC, performed well in classifying the data from all four datasets. Therefore, it was decided to choose three specific models for each data that showed good performance in general model training. Table IV shows validation accuracy for all subjects of the datasets, but we covered only ten subjects in NU data. This section is not considered in detail, and because of the time limit, we could not train each subject with cross-validation or ensemble learning.

V. CONCLUSION

In order to maximize the subject independence of the system, the suggested ensemble voting models, which are based on LDA, SVM, kNN, and LR classifiers, have been trained using generic training. For the ALS dataset, LDA and ensembe models produced the best test accuracy, which is 84.776%. The W-LDA-SVC-kNN model outperformed the other models for the ALS dataset in terms of f-score and variance. In terms of f-score value, our model outperformed [5]. For NU data, W-LDA-SVC-kNN demonstrated the best result in terms of variance. LDA slightly beat other models in terms of accuracy, although W-LDA-SVC-kNN for EPFLP data had the greatest f-score. It was expected that using ensemble model for classification of the datasets would drastically improve the accuracy and variance scores. Although, the ensemble models demonstrated better performance in some cases, the computational complexity of them is much higher. For further work, another types of machine learning algorithms with different hyperparameters will be used in combination to enhance the performance of ensembled models. Moreover, leave one out cross validation (LOOCV) is planned to be

implemented to increase the accuracy of the results

References

[1] Berdakh Abibullaev and Amin Zollanvari. "A Systematic Deep Learning Model Selection for P300-Based Brain–Computer Interfaces". In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 52.5 (2022), pp. 2744–2756. DOI: 10.1109/TSMC.2021.3051136.

[2] Sellers E Guger C Daban S. "How many people are able to control a P300-based brain-computer interface (BCI)?" In: *Neurosci Lett* 462.1 (2009), pp. 94–98. DOI: 10.1016/j.neulet.2009.06.045.

[3] et al Hoffmann U. "An efficient P300-based brain-computer interface for disabled subjects". In: *Neurosci Methods* 167.1 (2008), pp. 115–125. DOI: 10.1016/j.jneumeth.2007.03.005.

[4] Vinay Jayaram and Alexandre Barachant. "MOABB: trustworthy algorithm benchmarking for BCIs". In: *Journal of Neural Engineering* 15.6 (Sept. 2018), pp. 066–011. DOI: 10.1088/1741-2552/aadea0. URL: https://dx.doi.org/10.1088/1741-2552/aadea0.

[5] Ayana Mussabayeva, Prashant Kumar Jamwal, and Muhammad Tahir Akhtar. "Ensemble learning approach for subject-independent P300 speller". In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine; Biology Society (EMBC)* (2013). DOI: 10.1109/embc46164.2021.9629679.

[6] Angela Riccio et al. "Attention and P300-based BCI performance in people with amyotrophic lateral sclerosis". In: *Frontiers in human neuroscience* 7 (Nov. 2013), p. 732. DOI: 10.3389/fnhum.2013.00732.

VI. Contribution

Our team is very grateful for such an opportunity to greaten our knowledge in this field.

Aigerim was responsible for:

1) Introduction
2) Datasets info
3) Generic model training (with and without CV)
4) Exploratory Data Analysis with visualization
5) Subject-specific model training

Rasul was responsible for:

1) Abstract
2) Ensemble Learning
3) Results and Conclusion