Arousal-robust EEG Classification for Motor-imagery

Abdul Moeed
Technical University of Munich
Munich, Germany
abd.moeed@tum.de

***Abstract*—Developments in Brain Computer Interfaces (BCIs) are empowering those with severe physical afflictions through their use in assistive systems. Common methods of achieving this is via Motor-Imagery (MI), which maps brain signals to code for certain commands. Electroencephalogram (EEG) is preferred for recording brain signal data on account of it being non-invasive. Despite their potential utility, MI-BCI systems are yet confined to research labs. A major cause for this is lack of robustness of such systems. As hypothesized by two teams during Cybathlon 2016, a particular source of the system's vulnerability is the sharp change in the subject's state of emotional arousal. This work aims towards making MI-BCI systems resilient to such emotional perturbations. To do so, subjects are exposed to high and low arousal-inducing virtual reality (VR) environments before recording EEG data. The advent of COVID-19 compelled us to modify our methodology. Instead of training machine learning algorithms to classify emotional arousal, we opt for classifying subjects that serve as proxy for each state. Additionally, MI models are trained for each subject instead of each arousal state. As training subjects to use MI-BCI can be an arduous and time-consuming process, reducing this variability and increasing robustness can considerably accelerate the acceptance and adoption of assistive technologies powered by BCI.**

I. Introduction

Biological systems, such as humans, use electrical signals as the medium of communication between their control centers (brains) and motor organs (arms, legs). While this is taken for granted by most people, those with severe physical impairments, such as quadriplegia, experience the breakdown of this communication system rendering them unable to perform the most basic physical movements. Modern technologies, such as BCIs, have attempted to ameliorate this through the use of brain signals as commands for assistive systems [39]. MI, a common paradigm for BCI control, requires the subject to simulate or imagine movement of the limbs on account of there being discernible differences in brain signals when moving different limbs [1]. Due to it being non-invasive and cost-effective, EEG is the method of choice for collecting data for such systems [1].

One of the many recent developments in the application of EEG-driven BCIs is the Cybathlon competition held every four years under the auspices of Eidgenössische Technische Hochschule Zürich (ETH Zurich) [45]. The competition involves physically challenged individuals completing routine tasks via assistive systems. One such task – the BCI race – has the participants (called pilots) control a virtual game character via brain signals only. Competing teams, who may hail from either academia or industry, are responsible for creating BCI systems and training their respective pilots. The goal of the Cybathlon is to push the state-of-the-art in BCI assistive systems, and accelerate its adoption in everyday lives of those who need it most.

For the 2020 edition of Cybathlon, a team from the Technische Universität München (TUM) called "CyberTUM" is amongst the competitors in the BCI race challenge. In order to achieve high scores in the competition, a major part of BCI development is the focus on robustness of the system i.e. minimizing the variability of the system for different sessions and environments. Lack of robustness, in fact, is an established concern in almost all BCI systems. Possible causes of the problem include nonstationarity of EEG signals (variance for the same subject) [59] [56]. An additional cause, as noted by participating teams in Cybathlon 2016, is the change in the subject's emotional state. During the race, As expected, a public event such as the BCI race, the pilots' stress stress levels increased. This is to be expected as a public event such as the BCI race can heighten stress. This change in the pilots' emotional state caused their respective BCI systems to perform sub-optimally.

The objective of this work is to mitigate this concern and develop MI systems that are robust to perturbations in the subject's emotional state, specifically to emotional arousal. In order to achieve this, we develop VR environments to induce high and low arousal in the subject before recording MI data. VR environments have been previously used along with EEG to prompt changes in emotional arousal [5]. Additionally, they have been used together with MI for treating Parkinson's disease [32]. To our knowledge, this is the first work where VR environments are used to increase robustness of MI-BCI systems. Subsequently, learning algorithms are trained, not only for MI but also for different arousal states. The idea is that during the BCI race, we first detect the pilot's emotional state of arousal, and choose the appropriate MI classifier. Due to COVID-19, many steps in the above mentioned outline had to be modified, the details of which are present as follows.

II. Related Work

A. Cybathlon 2016

The inaugural Cybathlon competition was held in 2016. After the competition, the competing teams published their methods for training the participants, amongst which were

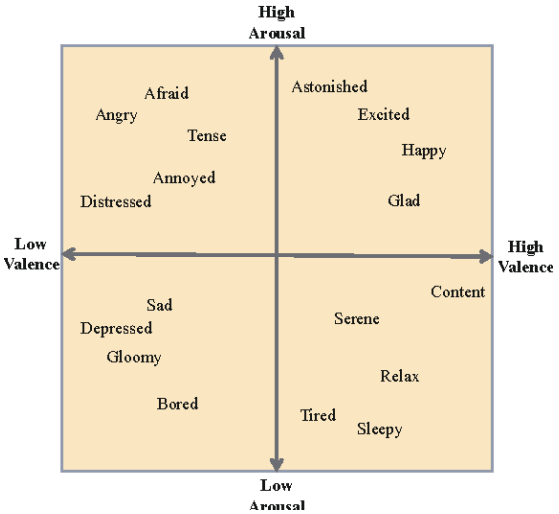


Fig. 1: The circumplex model of emotional classification. Figure courtesy of [19].

Brain Tweakers (EPFL) [39] and Mirage91 (Graz University of Technology). One of the pilots of the former performed well in the qualifiers but poorly in the final, prompting the authors to cite psychological factors such as stress as the possible cause for the drop. A similar course of events was observed for the pilot of Mirage91, who after achieving an average runtime of 120 s in the days leading up to the Cybathlon, dropped to 196 s during the competition. The authors indicated that the pilot was showing signs of nervousness on competition day, with a heart beat of 132 beats per minute (bpm) prior to the race [51].

The authors' hypothesis regarding the drop in their pilots' performances is supported by existing BCI literature [9] [28] [16] [18]. Further support comes from evidence in affective science: It has been theorized that any event that causes an increase in emotional arousal can affect perception and memory in a manner which causes the retention of high-priority information and disregard of low-priority information [31].

B. Emotional Valence and Arousal

Emotions are defined as complex psychological states, with three constituents: subjective experience, physiological and behavioral response [17]. Following early attempts [60], more rigorous descriptions of emotions were made, the most widely accepted of which being the 'circumplex model' [48]. It proposes that all emotions can be described as a combination of two properties: valence and arousal. These can be thought as orthogonal axes in two-dimensions. Neurologically, it entails that any emotional state is a result of two distinct and independent neural sub-systems [42]. Figure 1 provides a visual representation of the circumplex model. As can be seen, emotions such as 'excited' are high on both the arousal and valence axes, while 'gloomy' is low in both arousal and valence.

Alternative descriptions, such as the 'vector model' [7], do not veer off sharply from the circumplex model; they too base emotional classification on both valence and arousal. Hence the circumplex model was used as the paradigm of emotional analysis for the duration of the project.

C. Arousal, EEG and Motor Imagery

States of high and low arousal can be inferred from EEG signals [41]. This has been previously used to train learning systems for distinguishing between various arousal states [34]. EEG bands pertinent to different states of arousal are alpha (8-14 Hz) – related to a relaxed yet awakened state – and gamma (36-44 Hz) – a pattern associated with increased arousal and attention. The theta pattern (4-8 Hz), correlated with lethargy and sleepiness, is also useful for differentiating arousal.

With regards to motor imagery (MI), the most relevant EEG bands have been shown to be alpha (8-14 Hz) and beta (14-30 Hz) [15], the latter of which is associated with high degrees of cognitive activity [41].

Motor imagery data refers to data produced when the subject simulates limb movement. As movement of different limbs is sufficiently distinguishable, this can be used to perform control for various other tasks [37]. To record EEG data for motor imagery, the 10-20 international system of electrode placement is used 2. Due to the cross-lateral nature of limb control in the human brain, movement of the right arm is recorded most faithfully by C3 and that of the left arm by C4 [15].

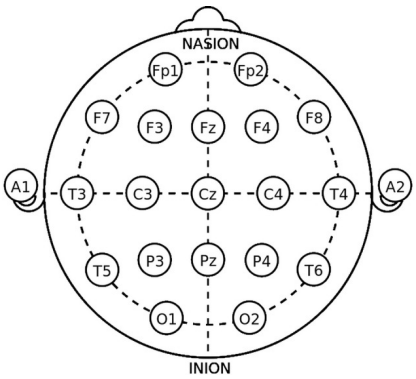


Fig. 2: 10-20 International system of EEG electrode placement. Electrodes C3 and C4 are most relevant for MI activity. Figure courtesy of [47].

III. Methodology

A. Virtual Reality Environments

Traditional methods of inducing stress include the Sing-a-song stress test (SSST) [8] and the Trier social stress test (TSST) [20], while meditation has been shown to induce relaxation [49] [29]. Emulating such environments faithfully in VR is sufficiently challenging, and may not be the most productive way to use VR to induce high/low emotional arousal.

Previously, VR exposure therapy has been explored to alleviate various psychological disorders [24]. One such example

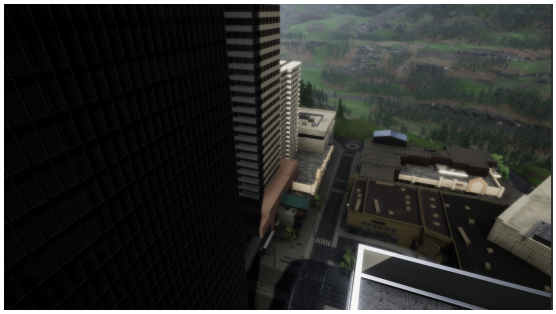


Fig. 3: Virtual reality environments for inducing arousal in subjects. On top is 'Height' designed to induce high arousal by placing the subject on the edge of a skyscraper. Below is 'Relaxation' intended to lower arousal via a natural, calming setting.

is using a VR height challenge – placing the subject on higher ground in a virtual environment [14]. Not only does the challenge induce high emotional arousal in test subjects, but the control subjects – the ones who are not acrophobic – also exhibit the same physiological responses as the test group i.e. increased heart rate and skin conductance level [14]. Similarly, VR environments, particularly those with natural scenery e.g. a forest, have shown efficacy in reducing stress [2] [3]. We thus developed two VR environments: one where the subject was placed on top of a skyscraper, called 'Height' while the second in a relaxing forest called 'Relaxation.' The environments were created using Unity 3D[1].

B. Dataset

As this project was part of the CyberTUM team's participation in Cybathlon 2020, the original idea was to collect real data with the actual pilots who will be competing in the even proper. At the beginning of this work, however, no ethics approval had been acquired to run any experiments on the pilots. This was not detrimental to the project as a proof-of-concept could still be arrived at by collecting EEG data from volunteers within the CyberTUM team. The COVID-19 pandemic obstructed our means of collecting such data.

In the absence of our own motor-imagery and arousal data, we opted for the Graz 2b data set [25]. It belongs to a family of BCI datasets collected by the BCI Lab at Graz University of Technology. The dataset has been used previously in the BCI Competition IV [53]. EEG data is collected for 9 subjects doing a binary motor-imagery task (moving right and left hand on cue). The data is sampled at a frequency of 250 Hz with 3 EEG and 3 EOG channels. For our experiments, we use data from two subjects, B05 and B04, whom we refer to as subject 1 and 2 respectively henceforth.

[1]https://unity.com/

C. Subject Classification as Proxy for Arousal Classification

As mentioned, we were unable to obtain our own EEG arousal data. To train the classifiers, we alternatively modified the experiment. Instead of using data with high/low arousal emotional states as labels, we used different subjects as proxies for such states, making it a cross-subject classification task [13] [46]. As EEG signals demonstrate significant variance between subjects, we can consider the data coming from subject A as that belonging to the emotional state of high arousal, and data from subject B as belonging to low arousal. With this approach, we can continue to train a classifier that would approximate the performance of one that is trained on actual arousal data, assuming the emotional states in this actual data are informative.

D. Experimental Design

The original scheme was to:

1) Develop VR environments in line with existing literature that are known to induce stress (high arousal) and relaxation (low arousal) in subjects.
2) Use electrodermal activity (EDA) activity to validate the efficacy of VR environments. EDA is a wide-used measure for emotional arousal, as skin conductance rises with rise in arousal [10].
3) Record MI data alternating between states of low and high arousal for each session. Start with 60s of inducing high arousal via the "Height" environment, then record MI data for 45s. Repeat the same with "Forest" environment for relaxed state. Repeat this process for each trial. The MI data was to be recorded by using the common paradigm of showing the participant a cue on screen (typically left or right arrow) which would prompt them to imagine as if they were moving their left or right hand [44] [40] [27].
4) Train an arousal classifier. The aim of this classifier is to indicate the emotional state (high or low arousal) of the subject.
5) Train separate MI classifiers for each emotional state. The goal is to optimize for accuracy, even if different types of pre-processing and classifier types were required for each state, unlike the arousal classifier which necessitates the same pre-processing steps.
6) During deployment, first classify the emotional state using the arousal classifier, and based on its result, choose the appropriate MI classifier.

As mentioned previously, due to numerous factors, many steps in the above formulation had to be either abandoned (2 and

3) or modified (4 and 5). The revised scheme, replaced steps 4-6 with the following:

1) Train a cross-subject classifier replacing the arousal classifier. The task of this classifier is to take EEG as input from any of the two subjects, and classify the input as belonging to either subject 1 or 2. As the classifier is agnostic to the subject, the same pre-processing had to be done for each subject's data.
2) Train separate MI classifiers for each subject instead of training for each emotional state.
3) At test time, sample a run of a few data points (5 in our experiments), feeding them to the cross-subject classifier. Based on its mode (most frequent classification), select the appropriate MI classifier.

E. Learning algorithms

We experimented with a multitude of machine learning algorithms which are briefly described as follows.

a) Logistic regression: Logistic regression is a modification of linear regression for a binary classification task [21]. It predicts the probability of a class given the input, by first learning a weighted linear combination of input features and applying a logistic function to the result.

$$y = \frac{1}{1+e^{-a}} \quad where \quad a = \theta_0 + \theta_1.x_1 + \theta_2.x_2 \qquad (1)$$

b) Linear discriminant analysis: LDA attempts to maximize inter-class variance while minimizing intra-class variance [4] in the data. This results in a clustering of the data where it is easily separable. It is widely used in MI BCI [58] [58].

c) Naive Bayes: A probabilistic classifier, naive bayes uses bayes' law to calculate the posterior probability of an event (class) given the prior and likelihood [33]. The posterior can then be updated with new evidence. It assumes that the features are independent, hence the term naive in its name.

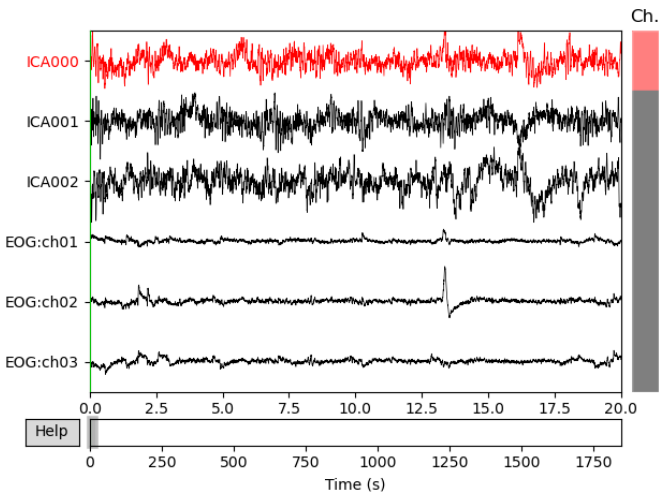
$$P(y|x) = \frac{P(y).P(x|y)}{P(x)} \qquad (2)$$

d) Ensemble model: This is implemented as a voting classifier in gumpy. It uses a mix of classifiers such as nearest-neighbor, LDA and support vector machines (SVM) and uses the majority vote as the classification output. As such, it necessarily either equals or outperforms both Naive Bayes and LDA as it uses them in the ensemble.

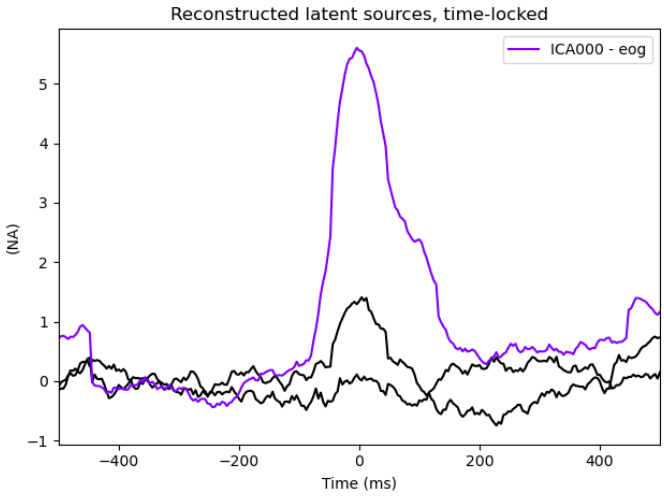
IV. Results

A. Artifact Removal

The data for subject 1 and 2 contained 324 and 399 trials (attempts at moving right or left hand) respectively. The standard approach to train MI classifiers is to analyze data and remove existing artifacts before extracting features from the data [55]. We first applied a Butterworth bandpass filter [11] to extract frequencies within the range 2-60 Hz. We then analyze the data for artifacts. A common source of artifacts in MI data is noise from electrodes located in the forehead's proximity. This is in fact data collected from the Electrooculography



(a) Plotting ICA with EOG channels. A visual depiction of the first component (in red) of ICA being correlated with EOG.



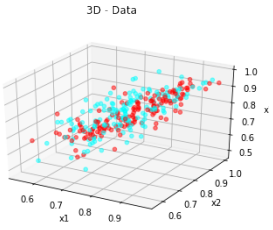
(b) Plotting ICA components against each other. The peak in the first component (blue) evidently due to an eye-blink.

Fig. 4: Artifact analysis using ICA for subject 1.

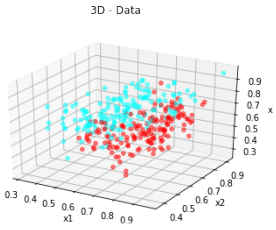
(EOG) channels which detect movements such as eye blinks, which may show up in the MI data. Such noise can be detected by first performing independent component analysis (ICA) – widely used in EEG preprocessing [30] – which tries to decompose a signal into constituent component under the assumption of statistical independence. We then see which of the resultant components correlates most with EOG channels, and filter it out [57]. An example of ICA on subject 1 can be seen in figure 4. We filter out the first component which seems to be picking up an eye blink. ICA on subject 2 did not improve the results.

B. Feature Extraction

Several methods were attempted to extract features. In principle, feature extraction in BCI takes two forms: frequency band selection and channel selection (also known as spatial



(a) PCA visualization of subject 1's feature vector.



(b) PCA visualization of subject 2's feature vector.

Fig. 5: Dimensionality reduction using PCA for feature space visualization of both subjects. Subject 2's features are more informative for the motor-imagery task compared to subject 1 which is also reflected in the training accuracy. Right hand movements are labeled red while left hand movements are blue.

filtering). In regards to the former, we've previously mentioned in II-C that alpha and beta bands have been shown to be most related to MI activity. Accordingly, we use these frequency bands as our features. In the same section we observed that channels C3 and C4 are the most relevant for MI, which we can use directly without any spatial filtering. For this, instead of using raw alpha and beta patterns, we opt for logarithmic sub-band powers of said patterns (see gumpy documentation[2]). Each spectrum is divided into four sub-bands. An alternative approach for feature extraction in MI classification has been the use of the "common spatial pattern (CSP)" algorithm [22]. It tries to find optimal variances of subcomponents of a signal [43] with respect to a given task. In our experiments, however, CSP performed poorly compared to logarithmic sub-band power of alpha and beta bands. The results when CSP was applied have thus been omitted from the report, but could be reproduced in the notebook (see section IX-A). A visualization of the features using PCA for both subjects can be seen in figure 5. As can be observed, the features for subject 2 are more conducive to discrimination of MI. This is also verified in the training results, where every classification algorithm achieved higher accuracy for subject 2 compared to subject 1.

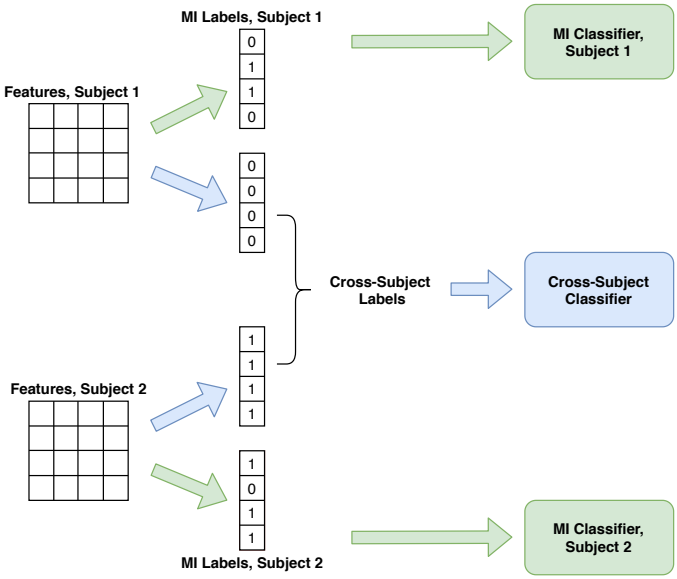


Fig. 6: Training scheme for both classifiers. MI classifiers are trained separately for each subject (labels corresponding to right and left hand) while Cross-subject classifier trained on features of both subjects.

C. Training

As mentioned previously, we train two types of classifiers: MI per subject classifier and cross-subject classifier. The entire training procedure is visually depicted in figure 6. After doing feature extraction, we first train an MI classifier for each subject with labels 0 and 1 (left and right hand movement respectively). Subsequently, we combine data of both subjects, labelling it 0 and 1 (subject 1 and subject 2 respectively) and train the cross-subject classifier. All classifiers described in III-E are trained in each case, the results of which can be seen in table I.

a) MI classification: The data for each subject was divided into an 80-20 split (training-test). The features were also standardized by rescaling to zero mean and unit standard deviation. Results for both subjects were satisfactory, although subject 1's data was harder to train on compared to subject 2. This can be observed by looking at the ranges of training accuracy for both subjects [55.84-70.12 vs. 91.25-95]%. Subject 2's classifiers achieved both a higher average accuracy as well as lower variance. LDA performed best for subject 1, while logistic regression achieved best results for subject 2.

b) Cross-subject classification: Training for cross-subject classifiers followed the same procedure of feature extraction with the only difference being a re-labeling of the samples from limb movements to source subject. Once again, we split the data into 80-20 (train-test) portions, though this time the data is the combined samples from both subjects. For testing the classifiers, we split the test set further into sections containing five samples (trials) each. For each section, we take the mode (most frequent prediction) of the classifier which is considered the final result. For example, if our test data has 50 samples from each subject, we portion it into 20 sections (each subject with 10 sections). We then feed each section

[2]http://gumpy.org/

to the classifier and take the majority score for that section as the classifier's prediction. As can be seen, the ensemble model outperforms the rest of the algorithms by a considerable margin. In addition to this, we also created t-SNE embeddings of the features with 2 and 3 dimensions [26]. The results were not up to par and have thus been left out here (they can be reproduced via the notebook discussed in IX-A). More details can be found in V.

TABLE I: Summary of results. Accuracy scores for MI (both subjects) as well as cross-subject (X-sub) using various classifiers. Best results in bold

Task	Classifier			
	Logistic Regression	**LDA**	**Naive Bayes**	**Ensemble**
MI-sub 1	67.53%	**70.12%**	55.84%	70.12%
MI-sub 2	**95%**	91.25%	91.25%	93.75%
X-sub	58.65%	59.38%	59.38%	**68.75%**

V. Discussion

The results indicate that assuming different emotional states impart sufficient differences in EEG data, we can train classifiers that perform well above chance. Significant differences in the EEG signals between both subjects were observed during feature extraction and classification. This is not an uncommon phenomenon and has been documented in the literature [41]. Blankertz et. al show that after testing on 80 subjects, the average classifier accuracy of a binary task was 74.4% with a spread of 16.5% [6]. Our findings buttress this as the best models for subject 1 and 2 achieved 75.38% and 95% accuracy respectively. This variability generally chalked up to differences in the subjects' abilities for implicit learning [23], performance in early neurofeedback sessions [35] and attention spans [12].

According to Tangermann et. al, the best results on data set 2b were achieved using filter-bank CSP as a pre-processing step followed by a naive bayes classifier during Competition IV [53]. In our testing, however, vanilla CSP for feature extraction was sub-optimal. Naive bayes was also found trailing behind other classifiers as seen in table I. We thus observe that vanilla CSP is not as performant as log band-power in our experiments, while we did not perform any experiments with filter-bank CSP.

In regards to cross-subject classification, appreciable results have been achieved by using ICA for feature extraction [54] combined with a nearest-neighbor (NN) classifier. We verify the efficacy of ICA as a pre-processing step for feature extraction. Other approaches have shown PCA as an effective step for dimensionality reduction [38]. While we could not confirm this with PCA, using the more modern dimensionality reduction technique of t-SNE performed poorly in our experiments (tested using target dimensions 2 and 3). There is, however, recent evidence that using t-SNE in tandem with common dictionary learning may yield good results [36].

A. Limitations and Future Outlook

A primary limitation of this work is the lack of testing on actual subjects. While the system ensures acceptable performance on an existing dataset, we can not conclude much about its usefulness in the real-world. To make such assertions with a certain degree of confidence, we need to evaluate how quickly we can switch between various MI classifiers based on the predictions of the emotion (cross-subject) classifier. This is also true for calibration time at the start of each session; while we use five trials during testing and get well above-chance results, comprehensive and systematic verification of the system is in order if it is to be of any practical use.

In addition to alpha patterns, gamma bands are correlated with increased arousal [41], which may have carried a strong supervision signal for the classifier. Had we acquired EEG data for aroused and relaxed states of a subject, an emphasis on gamma bands would have been warranted. As such, in the present case, as we did not have data corresponding to high and low arousal, gamma patterns were assumed not to be informative.

Future work may also look at training classifiers for more than two subjects. While two subjects suffice for the purposes of this study, as the original task was the discrimination between two emotional states of arousal, it may be worth exploring how the cross-subject classifier would scale to additional classes. This may be interpreted as having to classify not only emotional arousal but also valence (positive or negative) which may have important ethical implications.

Most of the classifiers used in this project are classic algorithms, and were chosen for their still prevailing use in MI BCI. However, future work may also incorporate modern approaches such as deep neural networks for MI classification [52]. Deep learning could also be used to formulate our problem as that of multi-task learning for both arousal and MI classification [50]. In this manner we can replace training multiple classifiers with a single one which both classifies emotional arousal as well as motor imagery.

VI. Interdisciplinary Work

The nature of this project necessitated the undertaking of a multi-disciplinary approach, from understanding and systematizing human emotional arousal to developing algorithms for distinguishing both emotional states as well as motor function via EEG. Thus, this work borrows, incorporates and synthesizes elements from a number of disciplines including psychology (emotional arousal), neuroscience (EEG and motor-imagery), computer graphics (virtual reality environments) and artificial intelligence (machine learning for classification). Broadly, we can categorize psychology and neuroscience as brain sciences and computer graphics and artificial intelligence under the umbrella of informatics. Each of the two disciplines contributed unique methods and insights without which the project may not have come to fruition. The most valuable insight was the difficulty in training accurate machine learning algorithms for EEG. Although machine learning has become the dominant paradigm for classification tasks, this project

demonstrates that pre-processing of data (via techniques such as ICA and log power-band) is at least as important to the success of the system as the classifier (the results for other feature extractors can be reproduced in the provided notebook), and even after pre-processing, we have no guarantees of robust performance. Another key insight was the extent to which EEG patterns vary between different people, pointing to the difficulty of transfer learning in this domain.

VII. Conclusion

A major hurdle in the widespread and practical use of assistive systems based on MI-BCI is lack of reliability. While this can have many origins, an important source as identified by two Cybathlon teams in 2016 was related to shifts in the subject's state of emotional arousal. In this work, we present an end-to-end framework for inducing high/low arousal in subjects, collecting EEG data and train learning algorithms for robust MI classification. While COVID-19 enforced certain constraints on data acquisition, we were still able to develop a proof-of-concept for how emotion-robust MI-BCI systems could be trained. Our results indicate that if the training signal contains sufficient information i.e. each emotional state has a distinct enough EEG signature, we can successfully train systems that are robust to variance in emotional arousal. A thorough study, however, needs to be conducted to determine the practicality of such a system with respect to variables such as classifier switching times and calibration periods.

VIII. Acknowledgments

This project could not have been possible without the aid of Nicholas Berberich who provided constant and quality guidance on overall methodology, feature extraction and algorithms. Also worth gratitude are Matthijs Pals for his support in regards to MI data preprocessing and Svea Meyer for helping in the initial phase of the project as well as with explaining EEG terminology.

References

[1] Reza Abiri, Soheil Borhani, Eric W Sellers, Yang Jiang, and Xiaopeng Zhao. A comprehensive review of eeg-based brain–computer interface paradigms. *Journal of neural engineering*, 16(1):011001, 2019.

[2] Allison P Anderson, Michael D Mayer, Abigail M Fellows, Devin R Cowan, Mark T Hegel, and Jay C Buckey. Relaxation with immersive natural scenes presented using virtual reality. *Aerospace medicine and human performance*, 88(6):520–526, 2017.

[3] Matilda Annerstedt, Peter Jönsson, Mattias Wallergård, Gerd Johansson, Björn Karlson, Patrik Grahn, Åse Marie Hansen, and Peter Währborg. Inducing physiological stress recovery with sounds of nature in a virtual reality forest—results from a pilot study. *Physiology & behavior*, 118:240–250, 2013.

[4] Suresh Balakrishnama and Aravind Ganapathiraju. Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*, 18:1–8, 1998.

[5] Thomas Baumgartner, Lilian Valko, Michaela Esslen, and Lutz Jäncke. Neural correlate of spatial presence in an arousing and noninteractive virtual reality: an eeg and psychophysiology study. *CyberPsychology & Behavior*, 9(1):30–45, 2006.

[6] Benjamin Blankertz, Claudia Sannelli, Sebastian Halder, Eva M. Hammer, Andrea Kübler, Klaus-Robert Müller, Gabriel Curio, and Thorsten Dickhaus. Neurophysiological predictor of smr-based bci performance. *NeuroImage*, 51(4):1303 – 1309, 2010.

[7] Margaret M. Bradley, Mark K. Greenwald, Mark C. Petry, and Peter J. Lang. Remembering pictures: pleasure and arousal in memory. *Journal of experimental psychology. Learning, memory, and cognition*, 18 2:379–90, 1992.

[8] Anne-Marie Brouwer and Maarten A Hogervorst. A new paradigm to induce mental stress: the sing-a-song stress test (ssst). *Frontiers in neuroscience*, 8:224, 2014.

[9] Ujwal Chaudhary, Niels Birbaumer, and Ander Ramos-Murguialday. Brain–computer interfaces for communication and rehabilitation. *Nature Reviews Neurology*, 12(9):513, 2016.

[10] Hugo D Critchley. Electrodermal responses: what happens in the brain. *The Neuroscientist*, 8(2):132–142, 2002.

[11] SS Daud and R Sudirman. Butterworth bandpass and stationary wavelet transform filter comparison for electroencephalography signal. In *2015 6th international conference on intelligent systems, modelling and simulation*, pages 123–126. IEEE, 2015.

[12] I Daum, B Rockstroh, N Birbaumer, T Elbert, A Canavan, and W Lutzenberger. Behavioural treatment of slow cortical potentials in intractable epilepsy: neuropsychological predictors of outcome. *Journal of Neurology, Neurosurgery & Psychiatry*, 56(1):94–97, 1993.

[13] Marcos Del Pozo-Banos, Jesús B Alonso, Jaime R Ticay-Rivas, and Carlos M Travieso. Electroencephalogram subject identification: A review. *Expert Systems with Applications*, 41(15):6537–6554, 2014.

[14] Julia Diemer, Nora Lohkamp, Andreas Mühlberger, and Peter Zwanzger. Fear and physiological arousal during a virtual height challenge—effects in patients with acrophobia and healthy controls. *Journal of anxiety disorders*, 37:30–39, 2016.

[15] Bernhard Graimann, Brendan Z Allison, and Gert Pfurtscheller. *Brain-computer interfaces: Revolutionizing human-computer interaction*. Springer Science & Business Media, 2010.

[16] Eva Maria Hammer, Sebastian Halder, Benjamin Blankertz, Claudia Sannelli, Thorsten Dickhaus, Sonja Kleih, Klaus-Robert Müller, and Andrea Kübler. Psychological predictors of smr-bci performance. *Biological psychology*, 89(1):80–86, 2012.

[17] H. Hockenbury. *Discovering Psychology*. Worth Publishers, Incorporated, 2000.

[18] Camille Jeunet, Emilie Jahanpour, and Fabien Lotte. Why standard brain-computer interface (bci) training protocols should be changed: an experimental study. *Journal of neural engineering*, 13(3):036024, 2016.

[19] Hye-Rin Kim, Henry Kang, and In-Kwon Lee. Image recoloring with valence-arousal emotion model. *Comput. Graph. Forum*, 35:209–216, 2016.

[20] Clemens Kirschbaum, Karl-Martin Pirke, and Dirk H Hellhammer. The 'trier social stress test'–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28(1-2):76–81, 1993.

[21] David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. *Logistic regression*. Springer, 2002.

[22] Z. J. Koles, Michael S. Lazar, and S Z Zhou. Spatial patterns underlying population differences in the background eeg. *Brain Topography*, 2:275–284, 2005.

[23] Irene Daum Marcus Schugens Niels Birbaumer Boris Kotchoubey, Stephan Haisst. Learning and self-regulation of slow cortical potentials in older adults. *Experimental aging research*, 26(1):15–35, 2000.

[24] Merel Krijn, Paul MG Emmelkamp, Ragnar P Olafsson, and Roeline Biemond. Virtual reality exposure therapy of anxiety disorders: A review. *Clinical psychology review*, 24(3):259–281, 2004.

[25] R Leeb, C Brunner, G Müller-Putz, A Schlögl, and G Pfurtscheller. Bci competition 2008–graz data set b. *Graz University of Technology, Austria*, pages 1–6, 2008.

[26] Ming-ai Li, Xin-yong Luo, and Jin-fu Yang. Extracting the nonlinear features of motor imagery eeg using parametric t-sne. *Neurocomput.*, 218(C):371–381, December 2016.

[27] Yi-Hung Liu, Shiuan Huang, and Yi-De Huang. Motor imagery eeg classification for patients with amyotrophic lateral sclerosis using fractal dimension and fisher's criterion-based channel selection. *Sensors*, 17(7):1557, 2017.

[28] Fabien Lotte, Florian Larrue, and Christian Mühl. Flaws in current human training protocols for spontaneous brain-computer interfaces: lessons learned from instructional design. *Frontiers in human neuroscience*, 7:568, 2013.

[29] Anna-Lena Lumma, Bethany E Kok, and Tania Singer. Is meditation always relaxing? investigating heart rate, heart rate variability, experi-

enced effort and likeability during training of three types of meditation. *International Journal of Psychophysiology*, 97(1):38–45, 2015.
[30] Scott Makeig, Anthony J. Bell, Tzyy-Ping Jung, and Terrence J. Sejnowski. Independent component analysis of electroencephalographic data. In *Proceedings of the 8th International Conference on Neural Information Processing Systems*, NIPS'95, page 145–151, Cambridge, MA, USA, 1995. MIT Press.
[31] M Mather and MR Sutherland. The selective effects of emotional arousal on memory. *Psychol. Sci. Agenda*, 2012.
[32] Anat Mirelman, Inbal Maidan, and Judith E Deutsch. Virtual reality and motor imagery: promising tools for assessment and therapy in parkinson's disease. *Movement Disorders*, 28(11):1597–1608, 2013.
[33] Kevin P Murphy et al. Naive bayes classifiers. *University of British Columbia*, 18:60, 2006.
[34] Tamás Nagy, David Tellez, Ádám Divák, Emma Lógó, Máté Köles, and Balázs Hámornik. Predicting arousal with machine learning of eeg signals. In *2014 5th IEEE Conference on Cognitive Infocommunications (CogInfoCom)*, pages 137–140. IEEE, 2014.
[35] N Neumann and N Birbaumer. Predictors of successful self control during brain-computer communication. *Journal of Neurology, Neurosurgery & Psychiatry*, 74(8):1117–1121, 2003.
[36] Takashi Nishimoto, Hiroshi Higashi, Hiroshi Morioka, and Shin Ishii. Eeg-based personal identification method using unsupervised feature extraction and its robustness against intra-subject variability. *Journal of Neural Engineering*, 17(2):026007, 2020.
[37] Natasha Padfield, Jaime Zabalza, Huimin Zhao, Valentin Masero, and Jinchang Ren. Eeg-based brain-computer interfaces using motor-imagery: Techniques and challenges. *Sensors*, 19(6):1423, 2019.
[38] Ramaswamy Palaniappan and Danilo P Mandic. Energy of brain potentials evoked during visual stimulus: A new biometric? In *International Conference on Artificial Neural Networks*, pages 735–740. Springer, 2005.
[39] Serafeim Perdikis, Luca Tonin, Sareh Saeedi, Christoph Schneider, and José del R Millán. The cybathlon bci race: Successful longitudinal mutual learning with two tetraplegic users. *PLoS biology*, 16(5):e2003787, 2018.
[40] Gert Pfurtscheller, Ch Neuper, Doris Flotzinger, and Martin Pregenzer. Eeg-based discrimination between imagination of right and left hand movement. *Electroencephalography and clinical Neurophysiology*, 103(6):642–651, 1997.
[41] Diego A Pizzagalli et al. Electroencephalography and high-density electrophysiological source localization. *Handbook of psychophysiology*, 3:56–84, 2007.
[42] Jonathan Posner, James A Russell, and Bradley S Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3):715–734, 2005.
[43] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller. Optimal spatial filtering of single trial eeg during imagined hand movement. *IEEE Transactions on Rehabilitation Engineering*, 8(4):441–446, 2000.
[44] Herbert Ramoser, Johannes Muller-Gerking, and Gert Pfurtscheller. Optimal spatial filtering of single trial eeg during imagined hand movement. *IEEE transactions on rehabilitation engineering*, 8(4):441–446, 2000.
[45] Robert Riener and Linda J Seward. Cybathlon 2016. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2792–2794. IEEE, 2014.
[46] Mouad Riyad, Mohammed Khalil, and Abdellah Adib. Cross-subject eeg signal classification with deep neural networks applied to motor imagery. In *International Conference on Mobile, Secure, and Programmable Networking*, pages 124–139. Springer, 2019.
[47] Gonzalo M Rojas, Carolina Alvarez, Carlos E Montoya, María de la Iglesia-Vayá, Jaime E Cisternas, and Marcelo Gálvez. Study of resting-state functional connectivity networks using eeg electrodes position as seed. *Frontiers in neuroscience*, 12:235, 2018.
[48] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.
[49] Peter Sedlmeier, Juliane Eberth, Marcus Schwarz, Doreen Zimmermann, Frederik Haarig, Sonia Jaeger, and Sonja Kunze. The psychological effects of meditation: a meta-analysis. *Psychological bulletin*, 138(6):1139, 2012.
[50] Y. Song, D. Wang, K. Yue, N. Zheng, and Z. M. Shen. Eeg-based motor imagery classification with deep multi-task learning. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2019.
[51] Karina Statthaler, Andreas Schwarz, David Steyrl, Reinmar Kobler, Maria Katharina Höller, Julia Brandstetter, Lea Hehenberger, Marvin Bigga, and Gernot Müller-Putz. Cybathlon experiences of the graz bci racing team mirage91 in the brain-computer interface discipline. *Journal of neuroengineering and rehabilitation*, 14(1):129, 2017.
[52] Yousef Rezaei Tabar and Ugur Halici. A novel deep learning approach for classification of EEG motor imagery signals. *Journal of Neural Engineering*, 14(1):016003, nov 2016.
[53] Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J Miller, Gernot Mueller-Putz, et al. Review of the bci competition iv. *Frontiers in neuroscience*, 6:55, 2012.
[54] Preecha Tangkraingkij, Chidchanok Lursinsap, Siripun Sanguansintukul, and Tayard Desudchit. Selecting relevant eeg signal locations for personal identification problem using ica and neural network. In *2009 Eighth IEEE/ACIS International Conference on Computer and Information Science*, pages 616–621. IEEE, 2009.
[55] Jose Antonio Urigüen and Begoña Garcia-Zapirain. Eeg artifact removal—state-of-the-art and guidelines. *Journal of neural engineering*, 12(3):031001, 2015.
[56] Carmen Vidaurre and Benjamin Blankertz. Towards a cure for bci illiteracy. *Brain topography*, 23(2):194–198, 2010.
[57] G. Wang, C. Teng, K. Li, Z. Zhang, and X. Yan. The removal of eog artifacts from eeg signals using independent component analysis and multivariate empirical mode decomposition. *IEEE Journal of Biomedical and Health Informatics*, 20(5):1301–1308, 2016.
[58] Yijun Wang, Shangkai Gao, and Xiaornog Gao. Common spatial pattern method for channel selelction in motor imagery based brain-computer interface. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 5392–5395. IEEE, 2006.
[59] Jonathan R Wolpaw, Niels Birbaumer, Dennis J McFarland, Gert Pfurtscheller, and Theresa M Vaughan. Brain–computer interfaces for communication and control. *Clinical neurophysiology*, 113(6):767–791, 2002.
[60] Wilhelm Max Wundt and Charles Hubbard Judd. *Outlines of psychology*, volume 1. Scholarly Press, 1897.

IX. Appendices

A. Documentation

The VR environments developed for this project can be found on the following Google Drive links:

- Forest VR - https://tinyurl.com/y79dxk87
- Height VR - https://tinyurl.com/yaqrmveu

The environments were created using Unity version 2018.4.14f with the post-processing stack enabled. All code written for this project pertaining to training can be found on LRZ Gitlab using the following link: https://gitlab.lrz.de/cybertum/eeg-classification. The Jupyter notebook titled 'Classifier Demo' contains the crux of the code; the rest of the scripts were created for testing tools and techniques. To reproduce ICA plots, run 'classifier.py'. The data used in this project can be found at http://www.bbci.de/competition/iv/ under 'Data sets 2b'. Further information on reproducibility can be found in the repository's README.

B. Methodological Reflections

1) *What would you do differently next time in your experimental/technical setup?* A major challenge, compounded by COVID-19, was the collection of arousal data. We were originally planning to use the Biopac system (BIOPAC Systems, Inc., Goleta, California, United States) for measuring skin conductance. The procedure is involved and somewhat cumbersome, and could have

been substituted with a heart rate sensor. Heart rate measurement is also an indicator of emotional arousal levels, and has the added benefit of being present in a smart watch. This would have made data acquisition simpler. Another important step would have been to test the efficacy of VR environments in inducing high/low arousal in systematic trials under control. Although there already exists evidence that VR has been known to impart such states, to observe its effect on EEG data would have been desirable. Finally, while classic machine learning is still widely used in MI classification, we observed, at least in subject 1's case, the difficulty in training a robust classifier. Perhaps opting for deep learning, which has shown increased performance in other domains such as computer vision and natural language processing, may have been promising.

2) *What went really good/easier than expected?* The creation of VR environments was surprisingly simple due to vast community support for Unity 3D. Integrating VR controls in the environments was slightly more involved but went rather smoothly. Regarding training: Besides feature extraction, training the classifiers was fairly straightforward. This was in part due to previous experience working with machine learning and in part because of easy-to-use implementations of common algorithms in existing libraries.
3) *What kinds of skills would you need to learn in your studies to do such a project even better (machine learning, statistics, psychology, neuroscience)?* Taking a course on BCI prior to working on this project would have substantially accelerated development. Trying to get familiarized with methods and terminology of the field was a major part of the project, but it also entailed allocating time for such self-study which could have been dedicated to development had prior knowledge in this domain been acquired. Thus a course on neurofeedback and BCI would have made the project better.
4) *If you tried out an approach which you later on abandoned, you can report this here as well. Give reasons why you decided not to pursue this approach.* During the training of MI classifiers, initially I used data from certain sessions for each subject. This yielded scores with high variance, even for the same subject for different sessions. I abandoned this by taking data from all sessions for each subject and using that for training instead.