LEARNING DATA REPRESENTATION AND EMOTION ASSESSMENT FROM PHYSIOLOGICAL DATA

Extended Abstract

Miguel Alexandre Rodrigues Tomás dos Santos Joaquim

Supervisor(s): Prof. Ana Luísa Nobre Fred
Prof. Hugo Alexandre Teixeira Duarte Ferreira

ABSTRACT

Emotional responses play a key role in Human communication and interaction with new technologies. Aiming at deeper understanding of Human emotional states, and to build adaptive, more effective Human-machine interfaces, we explore cutting-edge machine learning algorithms to analyse physiological data.

In a novel approach, two-channel pre-frontal raw electroencephalography and photoplethysmography signals of 25 subjects were collected using EMOTAI's headband while watching commercials. Taking as input the raw data, Convolutional Neural Networks were used to learn informative data representation and classify the acquired signals according to the Positive and Negative Affect Schedule. This unique approach achieved promising results, with average F1-scores of 76.6% for Positive Affect, and 83.3% for Negative Affect. Interpretation of the learned data representation was attempted by computing correlation values between various features extracted from the raw inputs and the final classification.

The same classification task was implemented using Support-Vector Machines and manually extracted features as input. The best average F1-scores, 79.0% for Positive Affect and 81.9% for Negative Affect were obtained with radial-basis function kernel and soft-normalization of the input features. Forward feature selection was used to check the features leading to the highest performance. Either way, the most important features appeared to be the alpha band power, and the asymmetry and phase synchronisation indexes. Therefore, the considered features seem to match the ones apprehended by the Neural Networks and selected by the Support-Vector Machines, hence endorsing their validity for emotional studies.

Lastly, biometric identification was also tried with Support-Vector Machines, achieving an average accuracy of 79.0%.

Keywords: *Emotions, machine learning, electroencephalography, photoplethysmography, Positive and Negative Affect Schedule, feature extraction;*

1. INTRODUCTION

For many years, the preferred method to evaluate someone's emotional state was self-assessment questionnaires. However, due to the rising interest in Brain-Computer Interfaces (BCI), this approach is becoming deprecated in favour of computational algorithms capable of assessing emotional states on their own. Furthermore, it opens the possibility to create digital interfaces which adapt themselves as to provide more intuitive and pleasant experiences to users based on their emotional responses.

Since the external stimuli which are responsible for eliciting emotional reactions also evoke changes in the central and peripheral nervous systems, it means that both are connected [1]. Consequently, different affective states are generally classified according to analyses of physiological patterns.

Emotion assessment based on physiological signals are usually implemented through machine learning algorithms, with 2 major data representation approaches. The first one finds patterns in physiological signals while using manually extracted features from already reported relations between patterns in physiological signals and emotional responses. Those features are used as input and mapped to new representations. In this case, although the chosen features are known, there is no way to know if they are the optimal ones. The second approach takes the raw signals as input and uses algorithms capable of extracting features on their own for classification. In those cases, the features should be optimal but are often unknown due to explainability issues.

Therefore, we propose combining both approaches and apply deep convolutional neural networks (CNNs), which are known for their outstanding feature learning capabilities [2], to perform emotion assessment according to the Positive and Negative Affect Schedule (PANAS) while using electroencephalography (EEG) and photoplethysmography (PPG) signals as input. Subsequently, the learned data representation is inspected to gauge if the EEG features commonly reported in the literature, such as temporal, spectral and connectivity features, match the ones apprehended by the neural networks. Time-domain PPG features are also considered to validate emotional responses.

Additionally, we explore feature-based classification of emotions with Support-Vector Machines (SVMs) which are, presently, the most widely machine learning algorithm for this type of tasks [3]. The performance between both classifiers (SVMs and CNNs) is compared, and feature selection is also implemented to check which combination of input features leads to the best performance for the SVMs. Afterwards, the selected features are compared to those apprehended by the CNNs to check if there is any similarity between the features considered by both approaches.

In the end, the emotion assessment task is explored from another point-of-view and converted into a simple identification task where different subjects are identified based solely on their physiological patterns collected for various emotional states. Moreover, we will investigate if different emotional states can also affect the identification process.

2. RELATION TO PRIOR WORK

From all the physiological signals that have been correlated with emotional responses, electrodermal activity (EDA) [4] electrocardiography (ECG) [5], PPG [6] and EEG [7] have been the most widely studied ones.

Regarding ECG and PPG, these signals have already been deeply studied due their simplicity to collect and analyse, so the metrics associated with behavioural changes have been known for many years. These include both the heart rate (HR) and heart rate variability (HRV) measures, with the latter becoming increasingly more popular in the recent years [8] since it is now well-recognized as a health indicator.

However, one of most popular physiological signals considered in studies aiming to evaluate emotional states is, definitely, the EEG as it reflects brain activity which is responsible for emotional regulation. Although innumerous EEG features have been studied and successfully used in emotion classification tasks, the optimal ones still remain unknown. For example, *Becker et al.* [9] implemented a valence assessment task using SVMs and other various machine learning techniques based on manually extracted features. He explored both time-domain statistics [10], frequency-domain features such as the powers of the major EEG frequency bands (θ, α, β, γ) [11], [12] and connectivity features such as the phase synchronization index (PSI) [13], [14]. Another feature often reported as heavily correlated with emotional responses is the pre-frontal asymmetry index of the α band [13], [16] with other authors claiming that frontal α asymmetry does not correlate with emotional responses by itself and depends on additional factors such as personality traits [17].

More recently, *Almogbel et al.* [18] successfully implemented CNNs to evaluate cognitive states for different types of workload using raw EEG data from *Muse* [19] as input but did not explore the apprehended features. Finally, although it was not used for emotion assessment, *Schirrmeister et al.* [20] used CNNs with raw EEG data as input to classify which motor brain region was being activated for a given task and explored the learned data representation by adding white noise to different frequencies of the input and computing correlation coefficients between input and output data.

Consequently, our work proposes to adapt an approach that has already been successfully applied in another context [20] to the area of emotion assessment tasks and expand the already existing studies to investigate if the features apprehended by an unbiased learner such as CNNs match the ones commonly used in other studies.

3. METHODOLOGY

3.1. Emotional model

PANAS [21] was chosen as the emotional scale for this study. It contains two independent scales: Positive Affect (PA) corresponding to positive experiences and positivity in general, and Negative Affect (NA) which is associated with negative experiences and perceived stress. Both scales are evaluated using 10 items, each one being classified from 1 to 5 for PA and -1 to -5 for NA. The final scores for PA and NA are computed by summing the values from their respective items. Additionally, we divided each scale into two classes, low and high by dividing both the PA and the NA scale in half (10-30 and 30-50). A final score higher than 30 (or lower than -30) was considered high and lower than 30 (or higher than -30) was considered low.

3.2. Stimuli

As elicitation protocol, we opted by using film excerpts as they present several benefits when compared to other commonly used alternatives. They are straightforward to implement and have proved to be one of the best inducers of both negative and positive emotions [22]. There are sets of validated video stimuli available [23] but most of them come from famous movies which make them unsuitable as the novelty factor is crucial in emotion assessment tasks. Consequently, random television commercials were chosen to elicit diverse emotional contents. Commercials contain many benefits that make them suitable for emotional studies. Their duration is short, have a defined target emotion and contain both audio and visible cues. In total, 14 commercials capable of eliciting a wide range of emotions (neutral, tenderness, amusement, sadness, disgust, anger and fear) were used. Furthermore, there were two commercials for each emotional response. The average duration of each commercial was 2:30 minutes.

3.3. Data acquisition

In this study, raw EEG and raw PPG data from 25 subjects, both male and female, with ages between 19 and 30 years, were collected with a sampling rate of 100 Hz using EMOTAI's wearable headband [24] which contains 2 EEG pre-frontal channels (Fp1 and Fp2) and a PPG sensor. Each subject watched 7 commercials, one commercial per each emotion, in a randomized order and evaluated it according to the PANAS scale, with 15 seconds interval between commercials to fill the PANAS questionnaires.

3.4. Data processing

Collected data was filtered to remove noise and undesired artifacts by using band pass filters. In order to generate the samples for the CNNs, an overlapping sliding window was applied to the physiological data corresponding to each commercial. Each window had a duration of 15 seconds (1500 points at 100 Hz) and were collected with a 5 seconds stride; hence 10 seconds overlap with the previous window. Each window was labelled with the PANAS classification of the complete commercial. The chosen dimensions allow the detection of enough peaks for a good estimate of PPG features while also keeping the EEG as short as possible due to its high variability. Furthermore, all the windows have the same dimensions (3×1500) which makes them suitable for deep learning applications. The first row corresponds to PPG data, the second to Fp1 and the third to Fp2. All rows contained raw, non-normalized data.

Regarding the samples to be used as input in SVMs, various features were manually extracted from the windows containing physiological data and stored into arrays with (1×N) shape, where N is the number of extracted features to be used as input. Therefore, each window to be used for the CNNs also generated an array to be used as input for the SVMs.

3.5. Proposed features

In order to check the features apprehended by the CNNs and the features selected by the SVMs, features which have been successfully used in feature-based emotion classification tasks were extracted from the EEG, both in the time and frequency domains.

- **Time-domain features:**

  - *Phase Synchronisation Index (PSI)* of the θ (4-8 Hz), low α (8-10 Hz), high α (10-13 Hz), β (13-25 Hz), γ (25-40 Hz) waves and of the complete signal [9]:

$$PSI(m,n) = \left|\frac{1}{T}\sum_{t=1}^{T} exp\{j(\phi_m(t) - \phi_n(t))\}\right| \tag{1}$$

    where ϕ_{m} and ϕ_n are the instantaneous phases of the EEG signal calculated from the Hilbert transform at brain regions *m* and *n*, respectively;

  - *Standard statistics* [9]: minimum (min), maximum (max), median, mean and standard deviation (std);

- **Frequency-domain features:**

  - *Asymmetry Index (AsI)* of the θ, low α, high α, β and γ frequency bands [25]:

$$AsI = \frac{L - R}{L + R} \tag{2}$$

    where L is the power of Fp1 and R the power of Fp2;

  - *Powers* of the θ, low α, high α, β and γ frequency bands [9].

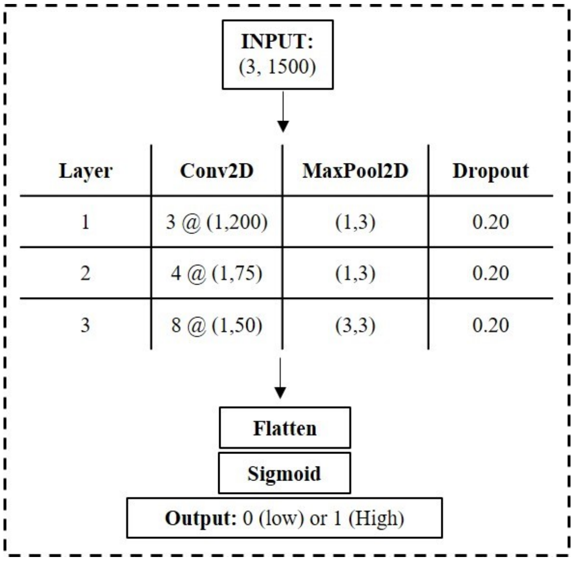
Besides the proposed EEG features, both the root mean square of successive differences (RMSSD) and HR were extracted from the PPG. Although they were not explored during the implementation and interpretation of the CNNs, both metrics were used to check if the commercials had been successful at eliciting adequate emotional responses and used as input features in the SVMs, together with the aforementioned EEG features. In order to check if different emotional responses were associated with statistically significant HR and RMSSD differences, the Mann-Whitney U test (**n** = 175; 7 commercials per subject, for 25 subjects) was applied. Frequency-domain features were not considered for the PPG analysis as they require, at least, 5 minutes recordings which was not the case [8].

3.6. Computational implementation

Since PA and NA are independent scales, one set of classifiers was trained per scale. The CNNs were implemented in Python using the Keras package with Tensorflow while the SVMs were implemented using the Scikit-learn package, also from Python. Both classifiers would output 1 or 0, corresponding to either high or low, respectively, PA or NA, depending on the scale being considered. Furthermore, for each type of classifier, two similar classification tasks, explained in 3.6.1 for CNNs and in 3.6.2 for SVMs, were implemented but with different validation approaches.

3.6.1. Convolutional Neural Networks

In the first task, an architecture that can be seen in Figure 1 was implemented. Validation and testing were done by randomly separating 15% of all the windows, for all the subjects, for testing and performing nested cross-validation on the remaining 85% of the data, with 15% of those samples being used for validation each time. The whole process was repeated 30 times. The performance of each classifier was evaluated by computing its accuracy and F1-score. The 3 best performing classifiers were then saved.



Layer	Conv2D	MaxPool2D	Dropout
1	3 @ (1,200)	(1,3)	0.20
2	4 @ (1,75)	(1,3)	0.20
3	8 @ (1,50)	(3,3)	0.20

Figure 1 Implemented CNN architecture

In order to interpret the data representation learned by the CNNs, white noise was added to the input raw EEG signals to change the proposed features in all the input windows. For the frequency-domain features, the signals were converted to the frequency domain, noise was added and then converted to the time-domain again. For the time-domain features, noise was added directly to the raw signals. The variation of the proposed features calculated at the input were calculated, the noisy signals were then classified, and variations at the layer right before the sigmoid function of the selected CNN were measured during classification. This allowed the computation of correlation's coefficients, and respective *P*-values between the features calculated at the input and output of the classifier. A single feature was considered each time and each procedure was repeated 250 times with the results of each run being the average result of the 3 saved CNNs. The final results were the average of the whole process and only the windows which kept their original classification (0 or 1) after noise was added were used. The false discovery rate (FDR) resulting from performing multiple comparisons was controlled using the Benjamini-Hochberg procedure [26].

However, in this first task, the samples from training and test sets could be heavily correlated due to the overlap of the windows containing physiological data, which limited the validity of the results. Consequently, we validated the obtained results with the second task. It was very similar to the first task except a leave-one-commercial-out approach was used. In this second task, all the windows from one random commercial per person were placed in the test set, meaning that all the windows containing physiological data in the test set were not correlated with the windows in the training set. Moreover, in this second task, we took advantage of all the windows belonging to a single commercial being either on the training or test set, and classified each commercial into high or low, PA or NA based on a majority voting from the classification of each individual window belonging to that commercial.

3.6.2. Support Vector-Machines

For the first task, validation and testing were done by randomly separating 15% of the samples (windows) for testing and performing nested 5-fold cross-validation on the remaining 85% of the data. The whole process was repeated 30 times. The performance of each classifier was evaluated by computing its accuracy and F1-score. The whole procedure was repeated for all the most commonly used kernels, linear, polynomial, radial-basis function (RBF) and sigmoid. Additionally, since different features have different orders of magnitude and values of higher magnitudes tend to have more impact on the final classification than the others, three distinct normalization methods were tested: no normalization, hard normalization and soft normalization. No normalization, as the name indicates, implies no normalization whatsoever. Hard normalization consists in, for a given feature, assigning the value of 1 to the highest value it takes and 0 to the lowest, with all the other values falling in between according to equation 3:

$$y = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{3}$$

where x is the real value of the feature and y its normalized value. Soft normalization, also known as standardization, normalizes each feature according to the distribution of its values as shown by equation 4:

$$y = \frac{x - \mu}{s} \tag{4}$$

where μ is average value of the feature, s its standard deviation, x its original value and y the standardized value. After considering all the possible combinations of kernels and normalization methods, the combination of features which led to the best performances were selected using an adapted forward feature selection algorithm. The algorithm starts with zero features and, per iteration, adds the feature which maximizes the performance of the classifier until the performance decreases. However, in this work, new features were added at every iteration until all of them were used.

Sequential Forward Selection Pseudo Code [27]

```
1. Create an empty set: $Y_k = \{ø\},\ k = 0$.
2. Select best remaining feature:
   $x^+ = arg\ max_{x^+ \in Y_k}[J(Y_k + x^+)]$
3. If $J((Y_k + x^+) > J((Y_k)$
     a. Update$Y_{k+1} = Y_k + x^+$
     b. $k = k + 1$
     c. Go back to step 2.
```

Similar to the first task with CNNs, the samples from training and test sets could be heavily correlated due to the overlap of the windows containing physiological data. Consequently, the results obtained with SVMs were validated through the second task. Everything was equal to the first task except leave-one-out validation was used and only the kernels and normalization method which had yielded the best results in the previous task were considered.

3.7. Identification problem

After implementing the emotional assessment task, an additional task was explored that consisted in identifying a person directly from their physiological signals and to test which emotion allowed a better differentiation between individuals. The first step was assigning to each commercial and its respective windows the ID of the person (subject 1, subject 2, subject 3, ...) who had watched it as the tag for classification. Then, the classification task was solved using SVMs as the number of samples per tag was much lower when compared to the original emotion classification tasks. Ergo, all the proposed features were extracted from each window and placed into feature arrays, once again. The kernel and normalization method which had yielded the best results in the emotion assessment tasks were used and one classifier was trained for each emotion. When training the classifier for a given emotion, all the windows corresponding to that emotion were placed in the test set while all the others were placed in the training set.

At the end of the learning process, the removed trials were classified, and the subject associated with each commercial was determined by a majority voting based on the classification of its windows. The whole process was repeated for all the 7 emotions with accuracy being the chosen metric to evaluate the performance of the classifiers. Lastly, this problem was also treated from an identification\authentication problem point of view. Instead of choosing just 1 subject per commercial based on majority voting, other individuals could also be identified as owners of the commercial (the person associated with the physiological data recorded during a specific commercial) as long as the number of trials assigned to that person surpassed a predefined threshold. In other words, a subject X would be accepted as owner of a commercial if the number of windows assigned to that subject surpassed the total number of windows of the commercial multiplied by the threshold value.

Both the False Acceptance Rate (FAR) and the False Rejection Rate (FRR) [28] were used as metrics to evaluate the behavior of the classifiers as biometric recognition systems. The former is the probability of falsely assigning the commercial to someone it does not belong to while the latter is the probability of wrongly rejecting the true subject who watched it. Both metrics can be easily calculated using Equations 5 and 6:

$$FAR = \frac{\#\,Incorrect\ Acceptances}{\#\,Attempts} \tag{5}$$

$$FRR = \frac{\#\,Incorrect\ Rejections}{\#\,Attempts} \tag{6}$$

In both equations, #Attempts is the number of times an identification process occurred. Considering a task with N individuals, each one with 7 commercials, the total number of commercials will be N×7. Additionally, there will be N attempts of identification per commercial, 1 per person, which gives $N^2 \times 7$ attempts in total.

Table 1: Average accuracies and F1-scores obtained with CNNs for the 3 groups in both tasks, where **n** is the total number of windows in each group.

		Fp1 (**n** = 2714)		Fp2 (**n** = 1760)		Fp1 + Fp2 (**n** = 522)	
		PA	NA	PA	NA	PA	NA
First Task	Avg. Acc (%)	75.1	79.3	75.3	87.0	78.6	86.5
	Avg. F1-score (%)	75.4	81.0	75.7	84.6	78.6	84.3
Second Task	Avg. Acc (%)	68.5	73.1	74.7	86.7	-	-
	Avg. F1-score (%)	67.5	75.6	75.2	84.6	-	-

By calculating both FAR and FRR for various thresholds, it was also possible to trace a ROC curve between FAR and FRR. Various thresholds (0.20, 0.35, 0.50, 0.65, 0.80) were chosen in order to simulate a system where, in the beginning, it is easy to assign the commercials to the wrong subject (threshold = 0.20) but becomes progressively harder to the point where the threshold is so high (threshold = 0.80) that there is the possibility of the true owner being rejected.

4. EXPERIMENTAL RESULTS

Regarding the PPG features, the *P*-values from the Mann-Whitney U test to check if there were significantly different HR values between high and low, PA and NA, were .158 and .042, respectively. For the RMSSD, the *P*-values for PA and NA were .001 and .030, respectively. Considering a significance level of .050, the results show that the commercials successfully elicited different emotional responses, especially according to the RMSSD, as they were not so significant for HR.

4.1. Deep learning classification results

When creating the dataset for the classification tasks, the collected data had to be separated into 3 distinct groups. One of them contained data from 15 subjects in which only Fp1 data was suitable to be analysed. The second contained data from 10 subjects in which Fp2 data was suitable to be analysed and the third group contained data from 3 subjects in which both Fp1 and Fp2 data could be properly analysed. For the first task, 2 sets of CNNs were trained per group, 1 for PA and 1 for NA, performing a total of 6 sets of CNNs. For the second task, 2 sets of CNNs were also trained per group. However, since the second classification task was only used to validate the first, the Fp1 + Fp2 group was not considered as it consisted in a subgroup of Fp2.

In all groups (as seen in Table 1), the obtained accuracies were above chance (as, on average, 54% of the windows were labelled as low PA and 60% were labelled as low NA). To compensate for the class unbalance, class weights were used during learning and a higher weight was given to the underrepresented labels, (inversely proportional to their amount). The best results were always obtained for NA which is consistent with the NA scale being more sensitive to external stimuli than PA [23]. The Fp2 channel performed better for NA than Fp1 as the right brain hemisphere is more relevant for negative emotions than the left hemisphere [15]. Supposedly, Fp1 is more associated with positive emotions but that behaviour was not observed. The third group, Fp1 + Fp2 had the best performances for both scales, showing that considering both brain hemispheres is important for emotion assessment.

When using the leave-one-out validation, correlations between the samples in the training and testing sets were completely removed. This led to slightly lower performances as there was no chance of any physiological data in the windows belonging to the test set to appear in the training, unlike what happened for the first task. Still, the results were similar to the ones previously obtained and above chance, hence validating the aforementioned results.

Table 2: Average correlation coefficients (and respective *P*-values) for different frequency power bands and statistics. Results with *P* < .050 are in bold. Significant results after correction are also marked with *.

		Band Power				Statistics	
Group	PANAS	θ	Low α	High α	Β	Median	Std
Fp1	PA	**0.237** **(0.048)**	**0.276** **(0.012)**	**0.225** **(0.015)**	**0.201** **(0.045)**	**0.055*** **(0.011)**	**0.203*** **(0.019)**
	NA	0.021 (0.382)	0.016 (0.318)	0.038 (0.272)	**0.152** **(0.018)**	**-0.012*** **(0.020)**	**0.102*** **(0.020)**
Fp2	PA	**0.197** **(0.012)**	**0.172** **(0.048)**	**0.115** **(0.047)**	0.020 (0.450)	**0.012*** **(0.017)**	**-0.212*** **(0.009)**
	NA	**0.166** **(0.042)**	**0.178** **(0.042)**	**0.173** **(0.020)**	0.028 (0.465)	**0.030*** **(0.011)**	**-0.184*** **(0.002)**

Table 3: Average correlation coefficients (and respective *P*-values) for different asymmetry and phase synchronisation indexes. Results with *P* < .050 are in bold. Significant results after correction are also marked with *.

		AsI			PSI	
Group	PANAS	θ	Low α	High α	θ	All
Fp1 + Fp2	PA	**-0.155** **(0.027)**	**-0.113** **(0.018)**	-0.027 (0.061)	**0.161** **(0.021)**	**0.337*** **(0.001)**
	NA	-0.101 (0.159)	**-0.156** **(0.032)**	**-0.221** **(0.043)**	**0.165** **(0.016)**	**0.342*** **(0.001)**

Table 4: Average accuracies and F1-scores obtained with SVMs for the 3 groups in both tasks for soft normalized input data and for the best performing kernel; **n** is the total number of feature arrays in each group.

		Fp1 (**n** = 2714)		Fp2 (**n** = 1760)		Fp1 + Fp2 (**n** = 522)	
		PA	NA	PA	NA	PA	NA
First Task	Kernel	RBF	RBF	RBF	Poly	RBF	RBF
	Avg. Acc (%)	74.4	73.0	74.4	81.7	85.9	86.5
	Avg. F1-score (%)	75.4	76.2	75.5	82.3	86.2	87.3
Second Task	Kernel	RBF	RBF	RBF	Poly	-	-
	Avg. Acc (%)	75.8	72.9	71.0	73.6	-	-
	Avg. F1-score (%)	75.9	75.7	72.0	79.0	-	-

4.2. Features' correlations

Due to the lack of space, only the most significant features for the various groups are shown in Table 2. Since CNNs are non-linear models, the magnitude of the correlation coefficients are not that meaningful, yet they can be used to assess which features have been apprehended by the network [29]. By comparing the different magnitudes for different features and their respective *P*-values, it is possible to observe how changing the value of different features in the input data influences the final classification.

The low frequency bands, namely θ and α bands, were on the verge of significance. This shows their importance in both Fp1 and Fp2 for positive emotions meaning that higher and coherent activity in both brain hemispheres is associated with positive emotions [16]. For negative emotions, however, these bands were only close to significance in Fp2, meaning that negative emotions tend to show higher brain activity mostly in the right hemisphere. In this case it is possible to observe Fp1's preference for positive emotions and how negative emotions are associated with Fp2 [16]. Statistical features managed to achieve significance, but the correlations were either much lower in magnitude or on par with those of spectral features, meaning they are not so impactful on the classification process.

Finally, regarding the connectivity features shown in Table 3, it is possible to observe near significant negative AsI correlations between both brain hemispheres. This was more prominent for the α band in the NA scale, meaning that a higher pre-frontal brain activity on the right hemisphere is associated with negative experiences. This is consistent with the pre-frontal α activity described in the literature for negative emotions [16]. Smaller asymmetries were also detected for PA, but these can be explained by PA not evaluating exclusively positive feelings. Moreover, during strong emotional responses, in both PA and NA, the synchronisation levels between both brain hemispheres tends to increase in the θ band and increases significantly for the whole signal as shown by the positive correlations of the PSI.

4.3. Support-Vector Machines results

Once again, due to the lack of space, only the detailed results from the best performing kernels and normalization method from the first task are shown. When creating the arrays with features to be used as input for the SVMs, the differences between the groups being considered had to be taken into account. For the Fp1 and Fp2 groups, only the PPG features and EEG features which could be analysed based on a single EEG channel were extracted, performing a total of 12 features. For the Fp1+Fp2 group, PPG features, and both single channel and connectivity EEG features could be extracted leading to a total of 33 features for that particular group. The normalization method with the highest accuracies and F1-scores was the soft-normalization, with the obtained results, as shown in Table 4, rivalling those of the CNNs shown in Table 1. The kernel with the best F1-scores was RBF for all the cases except on one case where the polynomial kernel had a better F1-score.

Once again, the F1-scores were slightly higher for the NA scale which is consistent with the behavior observed with the CNNs where the performances for the NA scale were consistently higher than those for PA. Similarly, Fp2 also obtained better results, confirming the importance of the right brain hemisphere during negative experiences compared to Fp1. Additionally, the best performances were obtained when using both Fp1 and Fp2 simultaneously which reinforces the importance of considering both brain hemispheres during emotion assessment tasks as it has also been observed with the CNNs.

Overall, it was possible to conclude that scaling and kernel choices play a major role in the performance of the classifiers. Furthermore, the SVMs with soft scaling out-performed the CNNs in the group Fp1 + Fp2. This may be due to the fact that the Fp1 + Fp2 contained fewer samples than Fp1 or Fp2 which made it harder for the CNNs to find patterns in the data when compared to the SVMs which do not require such a high amount of data. For Fp1 and Fp2, the number of available samples was bigger, hence the similar performance levels, with a slight advantage for the CNNs.

Regarding the second classification task, the average performances were very similar to each other and just slightly below the ones obtained in the first task, also with the already discussed slight advantage for the NA scale. While this decrease may be explained by the removal of the correlations between the training and testing samples by using a leave-one-out validation, the classifiers in this second task were still considerably above chance, so both the first and second tasks were successful.

In the end, considering both the CNNs and SVMs, it is noticeable that the performances obtained in the first and second tasks were much closer with SVMs than CNNs. This confirms that the amount of available data for training and testing, and the correlations between them can affect the performance of different machine learning algorithms. Since SVMs do not need as much data for training, they still managed to achieve similar performances after implementing a task that considerably reduced the correlation between the data. However, for the CNNs, one of its main limitations was the low number of samples available but the correlations in the data compensated for it. When eliminating the correlations, it resulted in slightly worse performances. Nonetheless, both approaches were still successful in both tasks at assessing different emotional responses solely based on physiological data.

4.4. Features' selection

The combination of features which resulted in the best performance of the SVMs for each group and scale was obtained using the aforementioned Forward Feature Selection algorithm but only for the kernels and normalization method which had achieved the best F1-scores during the first classification task. Therefore, all the obtained results come from applying soft normalization and the RBF kernel. Although the best performing kernel for Fp2 – NA when using soft normalization was the polynomial kernel, the implemented algorithm was taking too much time to perform the computations, so the RBF kernel was used instead as both kernel had similar performances (78.4% F1-score for RBF and 82.3% for the polynomial kernel). Starting with the results for Fp1, PA scale, they can be seen in Figure 2.

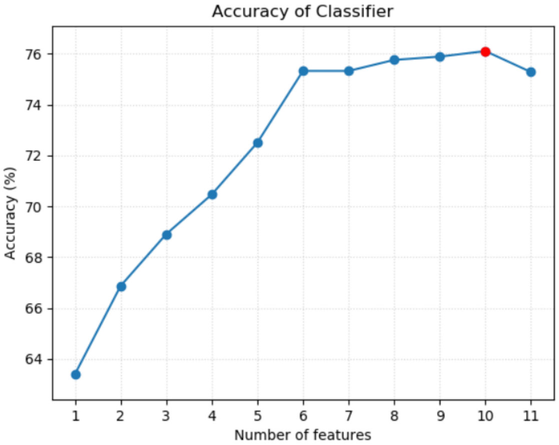


Figure 2 Average accuracies obtained while using forward feature selection for Fp1 in the PA scale.

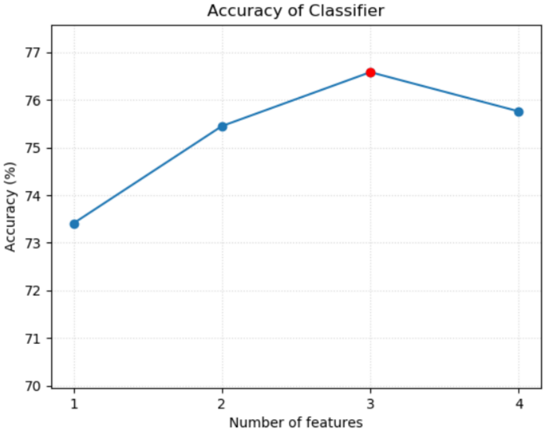


Figure 3 Average accuracies obtained while using forward feature selection for Fp1 in the NA scale.

A slight improve was observed with the average accuracy increasing from 74.4%, without feature selection to 76.1%, with feature selection. The selected features were the HR, RMSSD, power of θ, high α, β and γ bands, and the minimum, maximum, median and standard deviation of the EEG signal. Every available feature was selected with the exception of the power of the low α band and mean value of the EEG signal which means most considered features were relevant in the classification process. Comparing with the results from Table 2, five EEG features relevant in this situation were also considered relevant when using CNNs, which shows consistency in both approaches.

The results for Fp1, NA scale, can be seen in Figure 3. Another improvement was observed with the average accuracy increasing from 73.0% without feature selection to 76.6% with feature selection. In this case, the number of selected features was considerably lower, with only 3 features being needed to achieve the best performance. The selected features were the RMSSD, power of β band and standard deviation of the EEG signal which is also consistent with the results from Table 2.

When performing the same analysis for the Fp2 for the PA scale, as seen in Figure 4, the performance after feature selection was practically the same (74.8% after feature selection and 74.4% before feature selection).

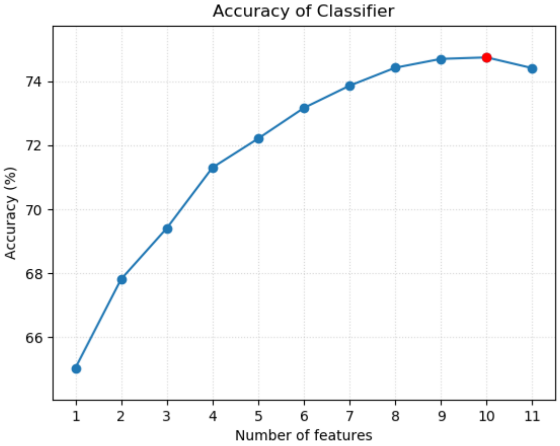


Figure 4 Average accuracies obtained while using forward feature selection for Fp2 in the PA scale.

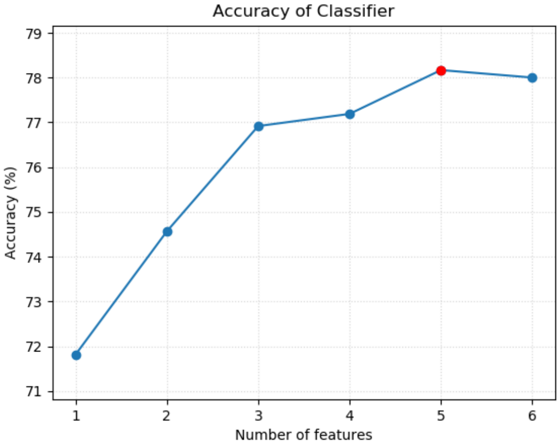


Figure 5 Average accuracies obtained while using forward feature selection for Fp2 in the NA scale.

The selected features were the HR, RMSSD, power of θ, low α, high α and β bands, and the minimum, maximum, median and standard deviation of the EEG signal. The best features were similar to those for Fp1 in the PA scale which supports the coherence between Fp1 and Fp2 for the regulation of positive emotions. Furthermore, when compared to the CNNs, 5 out of the 8 EEG features extracted with the SVMs were also considered as relevant for the CNNs, which shows some consistency between SVMs and CNNs, and Fp1 and Fp2 in the PA scale.

Similar to Fp1, when looking at the results for Fp2 for the NA scale (see Figure 5), the number of relevant features was considerably lower with only 5 features needed to achieve the best performance (74.6% before feature selection and 78.1% after feature selection).

In this case, the selected features were the RMSSD, power of *θ*, *low α* and high *α* bands and the mean of the EEG signals which is not only consistent with the results from the CNNs, but also reinforces the importance of the α bands in the right brain hemisphere for negative emotions.

The results of performing feature selection for Fp1+Fp2 for the PA scale can be seen in Figure 6, with the performance after feature selection increasing considerably (88.6% after feature selection and 85.9% before feature selection).

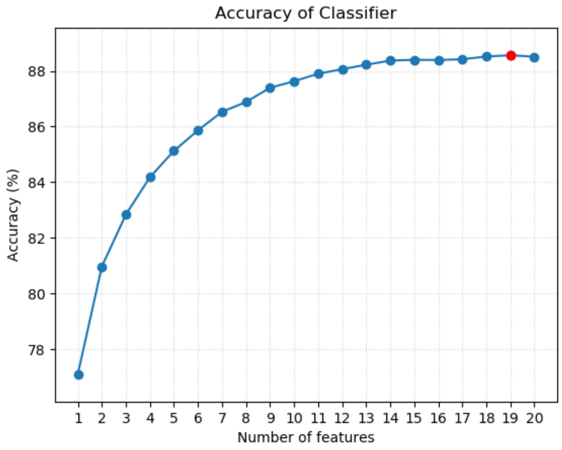


Figure 6 Average accuracies obtained while using forward feature selection for Fp1+Fp2 in the PA scale.

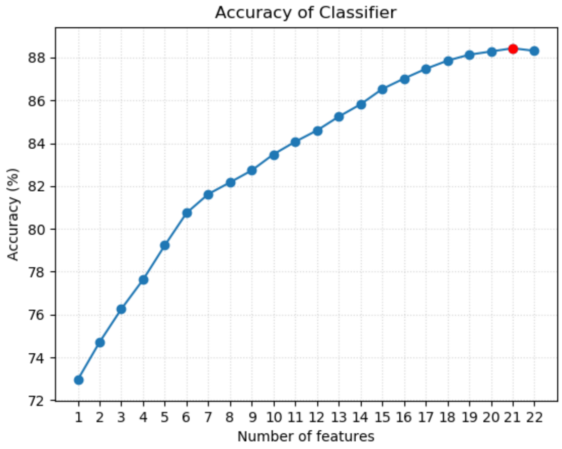


Figure 7 Average accuracies obtained while using forward feature selection for Fp1+Fp2 in the NA scale.

From a total of 33 possible features, 19 of them were selected to obtain the highest accuracy. From those 19 selected features, 8 of them corresponded to connectivity features (4 for AsI and 4 for PSI). This supports the role of considering both Fp1 and Fp2 simultaneously for emotion assessment. For the EEG features coming from only one channel, 6 of them belonged to Fp1 and 3 to Fp2. This shows a clear preference for Fp1 when analysing positive emotions. The last 2 features corresponded to the PPG features, HR and RMSSD.

Finally, the results of performing feature selection for Fp1+Fp2 in the NA scale can be seen in Figure 7, with the performance after feature selection also increasing slightly (88.4% after feature selection and 86.5% before feature selection).

From the 21 extracted features, 2 of them corresponded to the PPG features, HR and RMSSD and 9 of them to connectivity EEG features (AsI and PSI). Similar to the results for Fp1+Fp2 in the PA scale, this supports the major role of the balance between Fp1 and Fp2 for emotion regulation. For the 10 remaining features, 3 of them belonged to Fp1 and 7 to Fp2 which shows a clear preference for Fp2 in negative emotions. In total, from the 19 EEG features extracted with the SVMs, 12 of them were also considered relevant for the CNNs.

In the end, when comparing the features apprehended by the CNNs and the features selected by the SVMs across the various classifiers, it is possible to observe that the relevant EEG features in both approaches were very similar to each other. In both cases, the most important features were θ and α bands' powers, the standard deviation of the EEG signals and the connectivity features. Furthermore, for every classifier, there was at least 1 PPG feature being selected, confirming that optimal performances occur when combining data from both the PPG and EEG. When analysing both EEG channels simultaneously, more features from Fp1 were considered in the PA scale while more features from Fp2 were considered in the NA scale which, once again, supports the idea that the left brain hemisphere correlates more with positive emotions while the right brain hemisphere correlates more with the negative ones.

4.5. Identification problem

Based on the results from the emotion assessment task, the identification task was implemented using SVMs with the RBF kernel and soft normalization. Furthermore, since the original dataset was divided into three distinct datasets, this task considered just the two major groups, Fp1 and Fp2 as Fp1+Fp2 was just a subset of Fp2. The obtained results can be seen in Table 5.

Table 5 Accuracies from identifying the owner of each commercial for each emotion and group (Fp1, Fp2).

Emotions	Accuracy (%)	
	Fp1	Fp2
Neutral	80.0	80.0
Amusement	80.0	70.0
Tenderness	67.0	90.0
Sadness	73.0	90.0
Disgust	73.0	90.0
Anger	53.0	90.0
Fear	80.0	90.0

Considering that the Fp1 group had 15 people and Fp2 had 10 people, the obtained results show that it was possible to successfully identify the owner of each commercial in both groups for all the emotions with the exception of anger in the Fp1 group.

Furthermore, it shows a significant improvement from the 6.67% (for Fp1) and 10% (for Fp2) chances of finding the right owner of each commercial with a random classification. The Fp2 group obtained considerably better accuracies than Fp1 which can be explained by the lower number of subjects in that group. As the number of subjects in 1 group increases, the chance of finding 2 people with similar physiological patterns increases too.

In the Fp1 group, the best emotions to distinguish between subjects were fear, neutral and amusement followed by sadness and disgust. Anger was the emotion with the lowest accuracy, meaning that it was difficult to distinguish between subjects when they are angry. However, in the Fp2 group, all the emotions could be used to identify different subjects, even anger.

These were the results for the first approach of the identification task which could only attribute each commercial to one subject based on majority voting. The second approach consisted in attributing the commercial to any person who surpassed a predetermined threshold. The ROC curves between FAR and FRR for that second approach were calculated using the various thresholds mentioned in 3.7 and the final results can be seen in both Figures 8 and 9:

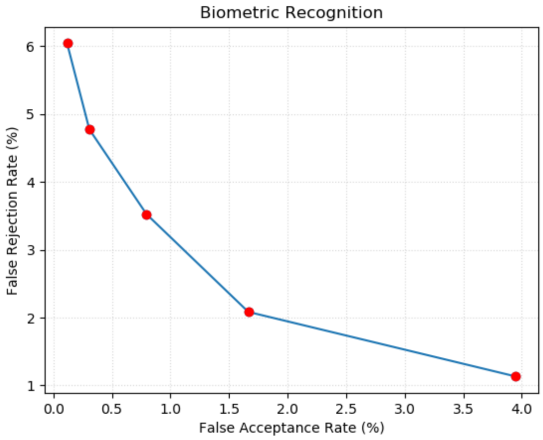


Figure 8 FRR/FAR ROC curve for Fp1.

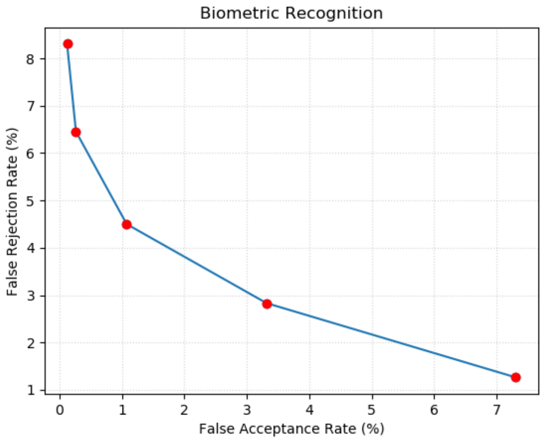


Figure 9 FRR/FAR ROC curve for Fp2.

In both cases, the FAR tended to increase as the threshold decreased while the FRR decreased, meaning that the chances of wrongfully accepting a subject as owner of the commercial increased. However, the chances the wrongfully rejecting the true owner of the commercial also decreased. Likewise, when the threshold increased, the FAR diminished and the FRR increased. It became progressively hard to be falsely identified as the owner of the commercial but the chance of rejecting the rightful owner also increased. Considering both groups and both metrics, the best threshold seems to be between 0.50 and 0.65 as that range seems to contain the closest point to both 0% FAR and 0% FRR.

In the end, creating a biometric recognition system based on PPG and EEG data would be a feasible approach and, depending on the type of application, the threshold could also be adjusted to make the system more loose or rigorous.

5. CONCLUSIONS

In this work, we proposed to use a combination of two-channel pre-frontal raw EEG and PPG signals as input for an emotion assessment task. This was successfully accomplished by using two different state-of-the-art machine learning techniques, Deep CNNs, which are capable of performing feature extraction on their own and SVMs using manually derived features. The obtained accuracies and F1-scores were considerably above chance which proves the benefit of considering both physiological signals simultaneously.

Interpretation of the learned data representation was attempted for the CNNs with the results being highly promising as the proposed EEG features seem to correspond to the ones apprehended by the CNNs. Furthermore, the proposed features were also used as input for the SVMs and feature selection was performed to verify which ones contributed most to the classification process. The selected features seem to be similar to the ones apprehended by the CNNs. Therefore, it is possible to conclude that variations in the proposed features can be tied to different emotional states.

Furthermore, through an adaptation of the classification algorithms, a biometric recognition system capable of identifying which subject matched each commercial for the 7 different emotions was demonstrated. It showed promising results reaching up to 90.0% accuracy for some emotions.

Future works in this continuously growing area should include more subjects and more recordings as results showed a clear pattern but most of them failed to achieve significance by a slim margin. Further studies should also look into additional methods of interpreting apprehended features, different neural network architectures and analyse other EEG features.

6. ACKNOWLEDGEMENT

This document was written and made publicly available as an institutional academic requirement and as a part of the evaluation of the MSc thesis in Biomedical Engineering of the author at Instituto Superior Técnico. The work described herein was performed at Institute of Biophysics and Biomedical Engineering, Faculty of Sciences, University of Lisbon (Lisbon, Portugal), during the period February-October 2019, under the supervision of Prof. Hugo A. Ferreira. The thesis was co-supervised at Instituto Superior Técnico by Prof. Ana L. Fred.

7. REFERENCES

[1] P. E. Ekman and R. J. Davidson, *The Nature of Emotion: Fundamental Questions*. Oxford University Press, 1994.

[2] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: a systematic review," *J. Neural Eng.*, 2019.

[3] S. M. Alarcao and M. J. Fonseca, "Emotions Recognition Using EEG Signals: A Survey," *IEEE Transactions on Affective Computing*, 2017.

[4] H. Sequeira, P. Hot, L. Silvert, and S. Delplanque, "Electrical autonomic correlates of emotion," *Int. J. Psychophysiol.*, vol. 71, no. 1, pp. 50–56, 2009.

[5] Z. Cheng, L. Shu, J. Xie, and C. L. P. Chen, "A novel ECG-based real-time detection method of negative emotions in wearable applications," in *2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC). IEEE,* 2018, pp. 296–301.

[6] R. Rakshit, R. V. Reddy, and P. Deshpande, "Emotion detection and recognition using HRV features derived from photoplethysmogram signals," in *ERM4CT 2016 - Proceedings of the 2nd Workshop on Emotion Representations and Modelling for Companion Systems*, 2016, p. 2.

[7] H. Zeng, C. Yang, G. Dai, F. Qin, J. Zhang, and W. Kong, "EEG classification of driver mental states by deep learning," *Cogn. Neurodyn.*, vol. 12, no. 6, pp. 597–606, 2018.

[8] F. Shaffer, R. McCraty, and C. L. Zerr, "A healthy heart is not a metronome: an integrative review of the heart's anatomy and heart rate variability," *Front. Psychol.*, vol. 5, p. 1040, 2014.

[9] H. Becker, J. Fleureau, P. Guillotel, F. Wendling, I. Merlet, and L. Albera, "Emotion recognition based on high-resolution EEG recordings and reconstructed brain sources," *IEEE Transactions on Affective Computing*, 2017.

[10] H. Xu and K. N. Plataniotis, "Affect recognition using EEG signal," in *2012 IEEE 14th International Workshop on Multimedia Signal Processing (MMSP). IEEE 2012 - Proceedings*, 2012, pp. 299–304.

[11] Z. Mohammadi, J. Frounchi, and M. Amiri, "Wavelet-based emotion recognition system using EEG signal," *Neural Comput. Appl.*, vol. 28, no. 8, pp. 1985–1990, 2017.

[12] W. L. Zheng and B. L. Lu, "Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks," *IEEE Trans. Auton. Ment. Dev.*, vol. 7, no. 3, pp. 162–175, 2015.

[13] Y. Y. Lee and S. Hsieh, "Classifying different emotional states by means of EEG based functional connectivity patterns," *PLoS One*, vol. 9, no. 4, p. e95415, 2014.

[14] M. Y. V. Bekkedal, J. Rossi, and J. Panksepp, "Human brain EEG indices of emotions: Delineating responses to affective vocalizations by measuring frontal theta event-related synchronization," *Neurosci. Biobehav. Rev.*, vol. 35, no. 9, pp. 1959–1970, 2011.

[15] D. Huang, C. Guan, K. K. Ang, H. Zhang, and Y. Pan, "Asymmetric Spatial Pattern for EEG-based emotion detection," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN), IEEE*, 2012, pp. 1–7.

[16] R. J. Davidson, "Anterior cerebral asymmetry and the nature of emotion," *Brain Cogn.*, vol. 20, no. 1, pp. 125–151, 1992.

[17] M. Palmiero and L. Piccardi, "Frontal EEG asymmetry of mood: A mini-review," *Front. Behav. Neurosci.*, vol. 11, p. 224, 2017.

[18] M. A. Almogbel, A. H. Dang, and W. Kameyama, "EEG-signals based cognitive workload detection of vehicle driver using deep learning," in *International Conference on Advanced Communication Technology, ICACT. IEEE*, 2018, pp. 256–259.

[19] "Muse." [Online]. Available: https://choosemuse.com/. [Accessed: 23-Oct-2019].

[20] R. T. Schirrmeister *et al.*, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapp.*, vol. 38, no. 11, pp. 5391–5420, 2017.

[21] J. R. Crawford and J. D. Henry, "The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample," *Br. J. Clin. Psychol.*, vol. 43, no. 3, pp. 245–265, 2004.

[22] R. Westermann, K. Spies, G. Stahl, and F. W. Hesse, "Relative effectiveness and validity of mood induction procedures: A meta-analysis," *Eur. J. Soc. Psychol.*, vol. 26, no. 4, pp. 557–580, 1996.

[23] A. Schaefer, F. Nils, P. Philippot, and X. Sanchez, "Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers," *Cogn. Emot.*, vol. 24, no. 7, pp. 1153–1172, 2010.

[24] "EMOTAI." [Online]. Available: https://emotai.tech/. [Accessed: 19-Oct-2019].

[25] M. K. Kim, M. Kim, E. Oh, and S. P. Kim, "A review on the computational methods for emotional state estimation from the human EEG," *Comput. Math. Methods Med.*, vol. 2013, 2013.

[26] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *J. R. Stat. Soc. Ser. B*, vol. 57, no. 1, pp. 289–300, 1995.

[27] A. Smith, O. Mendoza-Schrock, S. Kangas, M. Dierking, and A. Shaw, "An end-to-end vechicle classification pipeline using vibrometry data," *Ground/Air Multisens. Interoperability, Integr. Netw. Persistent ISR V*, vol. 9079, p. 90790O, 2014.

[28] "Biometric Recognition." [Online]. Available: https://www.bayometric.com/false-acceptance-rate-far-false-recognition-rate-frr/. [Accessed: 26-Jul-2019].

[29] G. Montavon, W. Samek, and K. R. Müller, "Methods for interpreting and understanding deep neural networks," *Digit. Signal Process. A Rev. J.*, vol. 73, pp. 1–15, 2018.