Interpretable Machine Learning Approach to Human Emotion Recognition and Visualization

Project Report for CS-GY 9223 Visualization for Machine Learning

Course Instructor: Dr. Claudio Silva

Vahan Babushkin[1*] , Binfang Ye[1*]

***Abstract*— Analyzing the biological signals our brain generates in response to external visual cues might shed the light on the elicited neurophysiological modifications in the nervous system that leads to the emergence of biological states known as emotions. Currently, there is no analytical model that can fully describe the processes in the human brain associated with emotions and can provide a reliable approach for reverse engineering of the biological signals into emotional states. However, several Machine Learning and Deep Learning technics are capable to infer emotions from biological signals with high accuracy. In comparison to the analytical model, the trained ML model or network acts as a black box, i. e. it does not provide enough information about the hidden processes in the human brain that are associated with the emotional state. The convolutional networks offer a prospective approach towards creating interpretable models for human emotion recognition. The feature maps generated in the hidden layers might shed a light on the biological signal features that can be associated with positive or negative emotions. In this project, we focus on the interpretability of the models for human emotion recognition from the electroencephalogram (EEG) recordings of neural activations elicited by visual stimuli. We conduct a visual analysis of the feature maps extracted by the hidden layers of two CNN models and conclude on frequency ranges that the model uses to differentiate between four emotion categories. Then we propose a modification of the currently existing model leading to the improvement of recognition accuracy. We also investigate the original data to eliminate outliers affecting the classification accuracy. Finally, we propose a pipeline for analysis of the EEG data for emotion classification.**

***Index Terms*— Emotion recognition, Affective computing, Visual Analytics, Machine Learning, Biomedical signal processing, Convolutional Neural Networks, Deep Learning**

I. Introduction

EMOTIONS play an essential role in our everyday lives since they affect our social interactions, decision making, perception and cognition, performance, and intelligence, and many other aspects of our lives. From the perspective of neurobiology, human emotions are associated with different biological states that originated as a response to neurophysiological modifications in the neural system elicited by external stimuli and are associated with thoughts, feelings, behavioral responses, and a degree of pleasure or displeasure [1]. From a psychological point of view the emotion consists of three components – arousal, affect, and feeling [2], which can be measured from the biological data. The psychological arousal of emotion is usually associated with the modified biological signals such as heart rate, skin conductance, and pupil dilation [3]. The behavioral demonstration of emotion (affect) is related to facial expression, gesticulation, or change in voice modulation. Conscious experience (feeling) of the emotion can be detected by analyzing changes in brain signals from electroencephalography (EEG) records [4]. Also, feelings can be estimated through self-reporting such as the self-assessment manikin [5]. However, these estimates are relatively subjective.

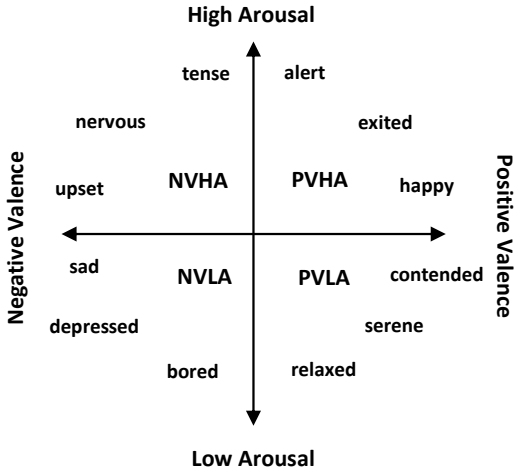


Fig. 1: 4 Emotion classes: Positive Valence High Arousal (PVHA), Negative Valence High Arousal (NVHA), Negative Valence Low Arousal (NVLA), Positive Valence Low Arousal (PVLA).

The concepts of valence and arousal act as dimensions in models for human emotions categorization. The commonly used circumplex model is demonstrated in Figure 1, where emotions are projected to the valence-arousal coordinates. The valence estimates the extent of the pleasantness of perceiving the stimulus and the arousal represents the degree of awareness induced by the stimulus. In most of the affective computing studies the valence and arousal are assigned high/low levels, thus distinguishing between positive valence high arousal (PVHA), positive valence low arousal (PVLA), negative valence low arousal (NVLA) and negative valence high arousal (NVHA) [6] [7] [8]. In this project, we follow the terminology in [3] where authors differentiate between Positive/Negative Valence (in most studies known

[1]V. Babushkin and B. Ye are with Tandon School of Engineering, New York University, NYC, 11201, USA vahan.babushkin@nyu.edu, binfang.ye@nyu.edu

*Both authors contributed equally to this work.

as High/Low Valence) and High/Low Arousal.

The accurate classification of human emotions is important in several applications where the intensity of emotion plays a crucial role, for example for human-machine interaction, mediated interpersonal communication, assistive technologies, and many others. To create algorithms that can recognize, process, mimic and influence human emotions, it is important to understand the biological mechanisms that are responsible for emotion generation. In order to recognize human emotions the AI system should rely on the biological signals, the human body produces in response to the external stimuli, that provide more objective insight into the human emotional state. For example, emotional markers can be found in electroencephalograms (EEG), and also can be deduced from heartbeat rates and skin conductivity. The EEG signals represent the neural activity in the human brain as a response to the external stimuli and thus related to the processes associated with the emotional state, while other modalities reflect the human emotional state indirectly [4]. This makes the EEG data a preferred source for human emotion classification. EEG signals can be measured immediately, and are not dependent on emotion-induced body responses such as changes in speech tonality or facial expression. In this work, we not only focus on classifying the human emotions from the EEG data, but also visualizing the model to gain insights that help us improve the model.

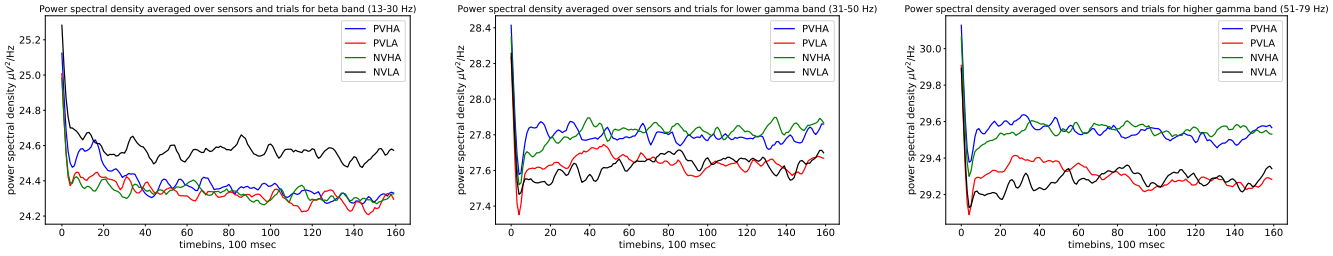
In this project, we use the EEG data for human emotion recognition according to the classical circumplex model (see Fig. 1). The main focus is on the interpretability of the models for human emotion recognition. We visualize the hidden layers of the Convolutional Neural Network (CNN) model described in [3] to study which features are responsible for the particular emotion category. It allows to make a conclusion about frequency ranges that the model uses to differentiate between four emotion categories. We propose a modified architecture of the CNN model that improves the recognition accuracy. A thorough analysis of original data determined the presence of outliers in spectral power density at the beginning of each trial, which affects the classification accuracy. These outliers were not reported in [3] and their removal increases the accuracy of the model. Finally, we propose a pipeline for analysis of the EEG data for emotion classification.

II. Related Work

The multiple studies demonstrate that it is possible to achieve high accuracy for human emotions classification using Machine Learning. Several factors affect the accuracy of emotion recognition such as different experiment environments, preprocessing techniques, feature selection, and length of the dataset [9] [10]. For example, on large datasets SVM usually achieves 33.3% and 25% accuracy for three and four emotion categories classification [10] compared to almost 80% accuracy with the same classifier reported in other studies. The popular Machine Learning approaches for human emotion recognition from EEG data include K-Nearest Neighbor (KNN), Bayesian Network (BN), Artificial Neural Network (ANN), and Support Vector Machine (SVM). For instance, the SVM model adapted for a multi-class classification was used for recognition of four music-induced emotional responses from EEG recordings with an accuracy of 90.52% using the one-against-one scheme and the accuracy of 82.37% with all-together scheme [11]. Other studies focusing on classifying the valence and arousal with SVM reported accuracies of 32% and 37% [12]. Usually, high accuracies are achieved while recognizing emotions with the SVM classifier from the offline data. However, for real-time emotion recognition, the classification accuracy is usually low. For example, in [13] the SVM classifier was used for online emotion recognition and achieved average accuracy of 70.5%.

The multivariate nature of the EEG data poses several challenges for emotion classification including noise and low generalization due to high dimensionality. Some studies use the Independent Component Analysis (ICA) for decomposing the data into independent components [4], or Empirical Mode Decomposition and then Genetic Algorithms to extract important statistical features [14]. Focusing on specific frequency bands and features increases the accuracy of the classifier significantly, e.g. applying SVM to five frequency features extracted from each channel of the EEG records leads to an average accuracy of 55.72% and 60.23% for classifying valence and arousal [15]. Incorporation of different modalities into the model such as audio/visual features, extracted from video stimulus increases the accuracy to 58.16% for valence and 61.35% for arousal [15].

The Convolutional Neural Networks (CNNs) gain popularity for emotion recognition from the EEG data in recent years due to their capability to simplify the data preprocessing stage and avoid engineering new features – the CNNs can learn the hidden dependencies in raw data by progressively encoding the features from primitive to more complex ones in subsequent layers. This property of CNNs enables detection of the local trends and extracts scale-invariant features for neighboring points, such as frequency variations patterns in nearby electrodes. It allows to capture the existing relationship between emotional states and the EEG data. For example, a CNN-based model proposed in [16] was able to achieve an accuracy of 85% for classifying valence and arousal, 77% for classifying positive, neutral, and negative valence/arousal, and 61% for classifying four categories of emotional states, namely low arousal/low valence (LALV), low arousal/high valence (LAHV), high arousal/low valence (HALV) and high arousal/high valence (HAHV). In [17] authors reported a 95.20% accuracy achieved with a CNN on DEAP dataset. And in the most recent study [3] the authors used SVM and CNN models to extract four classes of emotions (positive/negative valence high/low arousal combination) where SVM achieved an accuracy of 85% and CNN achieved 81%. They also consider the intensity of each of the 4 emotional categories by extracting 12 classes of emotional responses (low, medium, and high intensity for positive/negative valence high/low arousal combination) using the participants' self-report. The accuracies achieved



a.) b.) c.)

Fig. 2: The spectral densities for four emotion categories averaged over trials and channels. Notice that the magnitude of the emotion data does not strictly depend on the time except for the first 10 timebins ($\sim$ 500 msec).

on the 12 class dataset are 70% for SVM and 69% for CNN, which is quite high in comparison with the state-of-the-art emotion classification systems.

The drawback of CNNs is the need for large arrays of data for training which can be compensated by additional preprocessing of the input data or adding other modalities. For example in [18] authors use the EEG spectrogram and wavelet transformed galvanic skin response (GSR) to recognize the same four categories as in [16], achieving an accuracy of 73.43%. The EEG data can be processed separately for the theta (4-8 Hz), slow alpha (8-10 Hz), alpha (8-12 Hz), beta (12-30 Hz), and gamma (30+ Hz) frequency bands and consider the spectral power of all the symmetrical pairs of EEG electrodes within these five bands [19]. Ensembling of several classifiers, such as CNNs, Sparse Autoencoder (SAE), and Deep Neural Network (DNN) might also increase the accuracy up to 89.49% on valence and 92.86% on arousal recognition for DEAP and 96.77% for SEED datasets [20]. Other neural networks, such as ResNets for emotion classification from raw EEG data can achieve an accuracy of 90% for high/low valence and 58.03% for high/low arousal [21].

III. Methodology

We will be reusing the data of 34 subjects collected in [3]. For the experiment described in [3], 80 images were selected from the IAPS database that is known to elicit positive and negative emotions among humans. These pictures were categorized according to the circumplex model, i.e. positive valence high arousal (PVHA), positive valence low arousal (PVLA), negative valence low arousal (NVLA), and negative valence high arousal (NVHA). In total, 20 images for each class. The control of the experiment including the stimuli demonstration, stimuli synchronization with the experimental setup, and data collection was conducted using the Presentation software (by Neurobehavioral Systems, Albany, CA, USA) which collected the participants' response and event trigger information in the EEG system. The experimental setup consisted of a monitor to display the wash-off video and selected IAPS images to elicit emotions and a numerical keypad to collect rating responses from the participants. Participants were also asked to wear a pair of earphones with a white noise playing in the background to minimize the external auditory interference [3].

All 80 trials were breakdown into 4 runs (20 images each), representing the four classes of emotional responses (PVLA, PVHA, NVLA and NVHA, respectively). All runs were started/ended with a 20 seconds wash-off video to neutralize the emotional state of the participants'. Each image stayed on the display for a duration of 8 seconds before asking participants' for the rating on the keypad, followed by a 50 millisecond break period before loading the next image [3].

At the end of each trial the subjects were asked to rank their emotional state elicited by the visual stimulus on the scale from 0 to 9. This ranking allows further division of 4 emotional categories into 12 substates reflecting the extent of the subjects' emotional states [3].

IV. Data

According to the description in [3], a 1000 Hz sampled EEG signal was recorded through a 64-channel Brain Product EEG system. First, outside channels of FT9, FT10, TP9, and TP10 were removed, and then, a 0.1–85 Hz band-pass filter and 50 Hz notch filter, and a common average reference method were applied. After that, data was divided into two sets: the data set epoched according to the four stimulus events, and the data set epoched according to the rating of the participant's emotional status after viewing the stimuli. For the two data sets, the power spectral density of 1–80 Hz frequency bins was extracted through Short-time Fourier transform with 500 ms window was shifted by 50 ms. Each frequency bin had baseline correction using the one-second interval before image stimulation as a baseline. The processed EEG data represents power density from 2671 trials, each containing 160 timepoints recorded from 59 sensors and distributed across 80 frequency bins.

V. Visual Data Analysis

A. Spectral data visualization

We start by visualizing the spectral power densities to look for interesting patterns. For each emotion category, we average the data over trials and over all sensors grouping them over three frequency bands (Beta 13Hz – 30Hz, Low Gamma 31Hz – 50Hz, High Gamma 51Hz – 79Hz). The averaged plots of time course of spectral densities for each of the bands are shown in Fig. 2. The first observation is

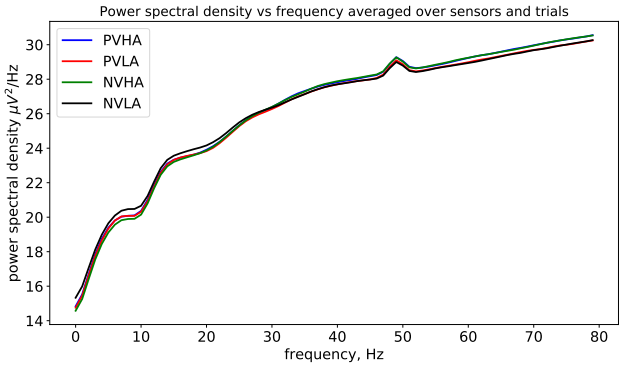


Fig. 3: Power spectral density averaged over 160 timepoints and 59 sensors.

that the magnitude of the spectral power densities remains relatively stable, except for the first 500 msec (10 time bins). It confirms that the emotion data is weakly dependent on the time within short time intervals (e.g. it takes minutes or even hours to alter the emotional state). We can also notice a significant increase in the spectral power density at the beginning on the trail, which can be explained either by the extensive firing of neurons on perceiving the new visual stimulus or, by the noise introduced with the trigger, which is a more plausible explanation. Further analysis suggests that this perturbation in spectral power density does not contain significant information about emotional states and its removal increases the accuracy of the model.

Another way to investigate the patterns in the spectral power is to consider for each emotion category the averages of spectral power densities over all trials, all sensors, and over time. The plots of spectral power versus all frequencies are demonstrated in Fig. 3. The spectral power increases with the frequency and demonstrates similar behavior for all four emotion categories. The peak at 50Hz reflects the effect of applying the notch filter. Interestingly, that the magnitude of the spectral power density for NVLA prevails for lower frequencies, particularly in the Beta band (13-31 Hz), while the magnitudes for NVHA and PVHA become more prevalent in Low Gamma (31Hz – 50Hz) and High Gamma (51Hz – 79Hz) bands. Overall, the power spectral density vs frequency plot demonstrates that the discriminating patterns for four emotion categories in the frequency domain are too close to each other which makes them hard to analyze.

B. Dimensionality Reduction

We performed the Principal Component Analysis with 3 components from the preprocessed dataset of 2671 trials averaged over time and three frequency bands (beta 13Hz – 30Hz, lower gamma 31Hz – 50Hz, higher gamma 51Hz – 79Hz) into the 2671×177 matrix (2671 trials, 177 features). The first three PCA components are shown on Fig. 4. The first component explains 51% of the variations in the data, the second 11%, and the last 7%.

We also applied T-Distributed Stochastic Neighbouring Entities (t-SNE) to the same data (see Fig. 5). The idea behind the t-SNE approach is to minimize the divergence between two distributions: one that represents pairwise similarities of the input objects and a distribution that represents the pairwise similarities of the corresponding low-dimensional points in the embedding. It allows to visualizes high-dimensional data by giving each datapoint a location on a 2D or 3D plane and helps to reveal the structure and patterns present in the data at many different scales [22]. For parameters, we used 3 components with perplexity (a measure of the effective number of neighbors) of 40. The algorithm achieved a mean sigma of 6.99 and KL divergence of 0.19 after 5000 iterations.

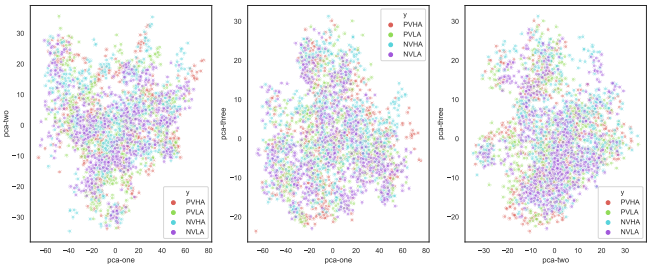


Fig. 4: Results of applying PCA to the data of 2671 trials collapsed into a vector of 177 features for each trial. The first component explains 51% of the variations in the data, the second 11% and the last 7%.

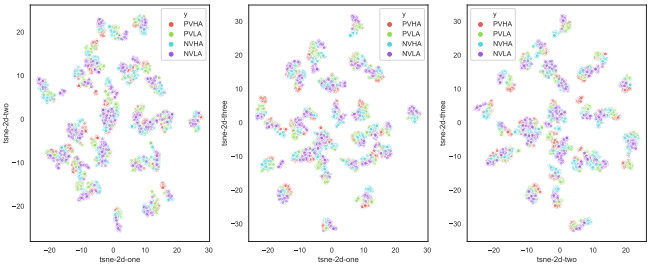
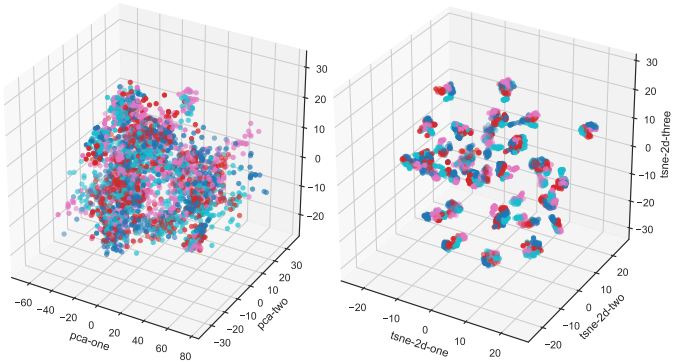


Fig. 5: Results of applying t-SNE with 3 components with perplexity of 40 to the data of 2671 trials collapsed into a vector of 177 features for each trial. The mean sigma was 6.99 and the KL divergence of 0.19 was achieved after 5000 iterations.



a.) b.)

Fig. 6: The 3D plots of first three components of PCA a.) and tSNE b.).

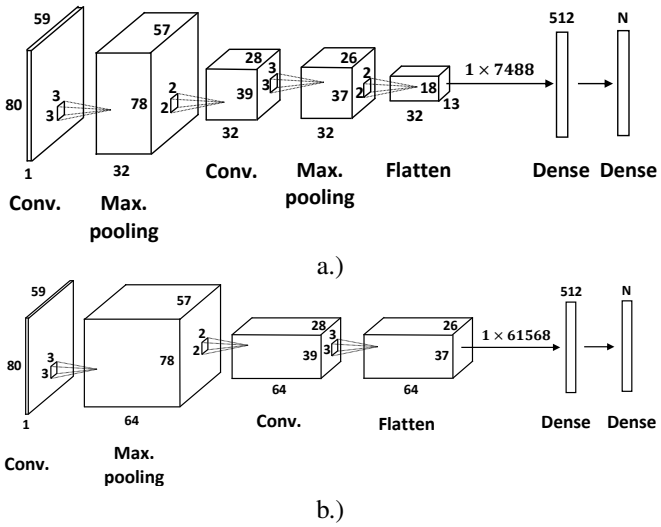


Fig. 7: CNN architectures for classifying 4 emotion categories: a.) used in [3], b.) our proposed architecture.

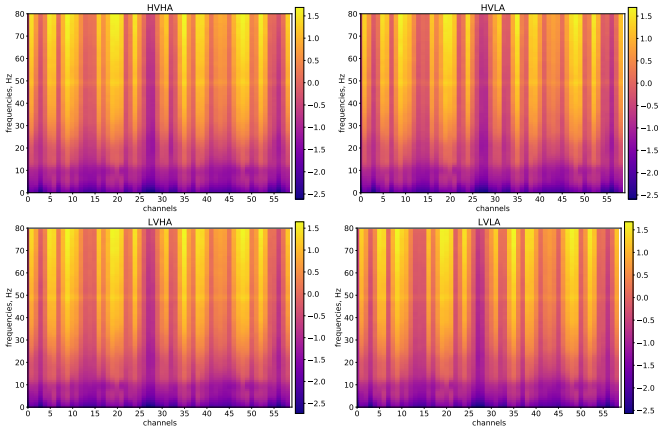


Fig. 8: Spectrogram of the four emotion classes

The 3D plots of PCA and t-SNE components are demonstrated on Fig. 6. While in the PCA plots (Fig. 4 and Fig. 6, a.)) it is hard to discern between clusters formed by each of emotion categories, the t-SNE plots (Fig. 5 and Fig. 6, b.)) show an interesting combination of several clusters each containing an agglomeration of clearly discernable four subclusters that correspond to four emotions. The formed clusters might be interpreted as intensity levels of emotional categories and subclusters – the four emotional categories. Thus, the visualization suggests that it might be incorrect to classify the multitude of human emotions just into four classes. It would be more relevant to consider different intensities of arousal and valence, eventually converging the problem of emotion classification into regression problem, where arousal and valence are qualified with either positive or negative continuous value. It will require redesigning of the experimental setup and additional data collection.

VI. Visualizing Hidden Layers of the CNN Network

First, we visualized the layers of the CNN model presented in [3] and shown on Fig. 7, a.). We use the same inputs as in [3], i.e. we averaged the 4D tensor of $80 \times 160 \times 59 \times 2671$ was across the second dimension (time) and added an additional singleton dimension to the newly-created 3-D tensor of $80 \times 59 \times 2671$ to make it suitable for processing with CNN. Later, based on the visualizations we have reviewed the architecture of the CNN network proposed in [3] by removing the last max-pooling layer before the dense layer and increasing the number of channels in convolutional layers to 64 (see Fig. 7, b.)).

We start with visualizing the feature maps produced by each layer of the old and the new model. First, we visualize the input spectrograms of four classes of test data averaged over all trials from each class (Fig. 8). At the first glance, they are not very informative in the sense that there are no visible patterns that differ across the classes. Then we plot the feature maps of the convolutional and max-pooling layers averaged for all classes of the CNN model from [3], Fig. 7, a.) to investigate what kind of information is extracted by the hidden layers (see Fig. 13 in Appendix). One observation is that the feature maps visually appear similar to spectrograms, with the bottom of the maps corresponding the lower frequencies and top – to higher. There is a persistent pattern across different features, particularly noticeable in later layers – some filters give more weights to specific frequency ranges between beta and high gamma (13Hz - 80Hz), others to the lower frequency bands as delta (0.1 to 3.5Hz), theta (4-8 Hz), and alpha (8-12Hz). Some filters look for specific frequency-dependent features for every sensor (sharp vertical lines), others extract common patterns for groups of "neighboring" sensors (smooth vertical lines). It appears that that the distribution of power spectral density across channels for a given frequency is the detrimental feature that allows to discriminate between emotion classes. We can also noticed that the feature maps extracted by different layers differ between instances of each class (Fig. 15 - Fig. 18) which is particularly noticeable in the later layers (Fig. 18). For example, for PVLA the more weights are given to the higher frequency bands, while for PVHA activated lower frequency ranges (delta, theta, and alpha) and high gamma prevail. Thus, the CNN is capable of extracting the feature maps that determine the power spectral density distribution in frequencies and sensors, characterizing each particular type of emotion.

Observing the decrease in resolution of feature maps after the last max-pooling layer and the complexity of patterns extracted by each of the filters, we decided to increase the number of filters and to remove the last max-pooling layer to provide the dense layer with more features (Fig. 7, b.)). The 10-fold cross-validation of the new model demonstrated an increase in the accuracy from 0.83 to 0.84 (see Table I). Similar visualizations can be conducted for the hidden layers of the new model. Fig. 14 shows the feature maps of the hidden layers of the proposed model, averaged over all four emotion categories. We also show the feature maps extracted by different layers for correctly classified instances

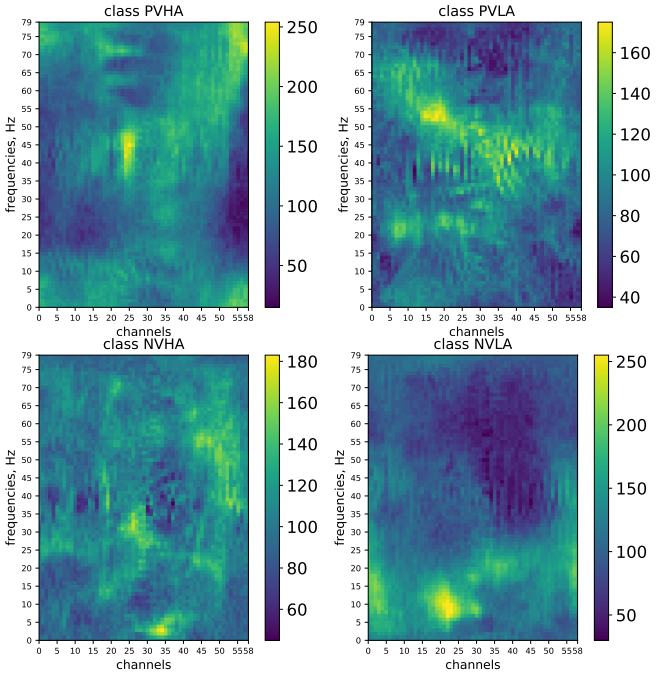


Fig. 9: Visualization of dense layers with activation maximization for each class of the model in [3] (Fig. 7, a.)).

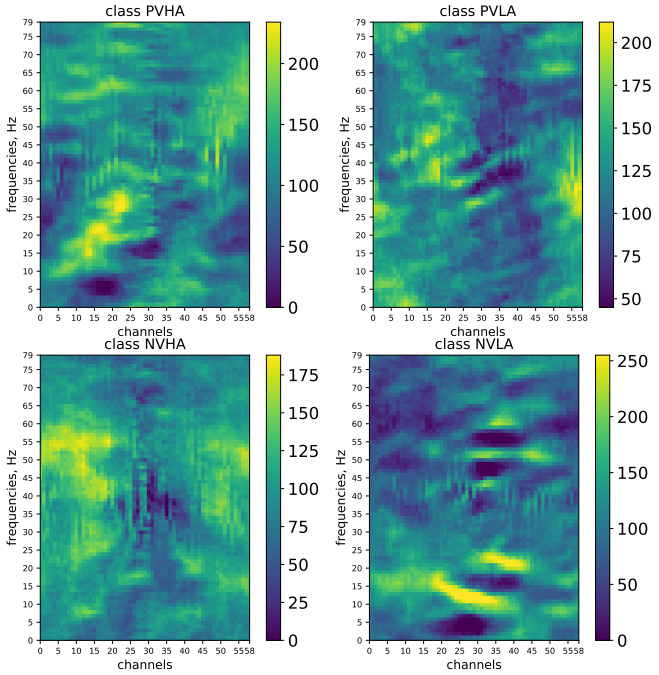


Fig. 10: Visualization of dense layers with activation maximization for each class of the proposed model (Fig. 7, b.)).

of each class (Fig. 19 - Fig. 20). Noticeably, the filters of the last convolutional layer produce feature maps, more clearly discernable among the categories (Fig. 20). We can also notice a parallel line at approximately 50 Hz in later layers in both architectures, which most probably represents the perturbation caused by 50Hz notch filter.

To explore which parts of the input spectrogram maximize the correct classification of emotions we applied an activation maximization technic to the dense layer for both models for 4 emotion categories with the help of tf-keras-vis visualization toolkit. The activation maximization synthesizes an artificial image that, if used as the input, would maximize the target response [23]. The activation maximization of the last dense layers for old and new models are demonstrated on Fig. 9 and Fig. 10 correspondingly. Visualization with activation maximization of the last layer shows that there is a clear difference between the four classes, which is more pronounced for the new model. In general, the NVLA is characterized by activations in alpha (8-12 Hz) and beta (13Hz – 30Hz) bands for channels from 15 to 45 while the NVHA can be described by the activations in lower gamma and in high gamma for channels between 0 and 20 and 45 and 59 (see Fig. 10). For PVLA we have the activated regions in lower/higher gamma – wider and blurred in the old model (from channels from 10 to 50) but sharper in the new model, where the lower gamma activations are observed in channels between 55 and 59 and lower/higher gamma – for channels between 10 and 30. For PVHA the activation mostly happens in beta for channels from 10 to 25 and in high gamma for channels 50-59, but the old model does not capture the beta band activations. These observations are in accordance with the visualizations of spectral power densities shown on Fig. 2. Indeed, from Fig. 2, a.) we can notice that in the beta band the magnitude of power spectral density over time prevails for NVLA, while in lower gamma and higher gamma bands the magnitudes for PVHA and NVHA are most prevalent. Given that the temporal resolution of EEG is very high, it is totally valid to claim that NVLA state is characterized by oscillations in the beta band, NVHA and PVLA – in lower gamma and in high gamma, and PVHA – beta and high gamma. However, it would be incorrect to associate the activations for different emotion categories with corresponding brain regions due to the very poor spatial resolution of the EEG. The best way to investigate this problem would be to redesign the experiment by combining a high temporal resolution technics, as EEG with a high spatial resolution one as fMRI. Performing the source reconstruction from EEG data might also shed light on the locus of emotion in the human brain, however, it will require processing massive amounts of data.

The relative obscurity of visualized activations in an old model was another objective for reviewing the model's architecture. The removal of the max-pooling layer in the old model on Fig. 7, a.) results in more clear activation maximization visualizations for the last dense layer. In addition, increasing the model width, i. e. the number of filters in each convolution layer from 32 to 64 leads to the improvement of the classification accuracy.

VII. Visualizing SVM Model

We were also interested in investigating how the SVM model, proposed in [3] is capable of discerning between 4 types of emotions. For this purpose, we visualized the decision boundary drawn by the SVM model with RBF function with $\gamma = 0.1$ for the $80 \times 160 \times 59 \times 2671$ tensor averaged across the time and three frequency bins,

TABLE I: Evaluation of old and new models on the data with the first 10 time bins ($\sim$ 500 msec) trimmed and low frequencies (from 1 to 13Hz) removed. The performance metrics are obtained by 10-fold cross-validation of the models, and averaging over all folds. For each fold the model has been trained for 100 epochs.

Metrics	Old model			New model		
	Full Data	Time Trimmed	Time & Freq. Trimmed	Full Data	Time Trimmed	Time & Freq. Trimmed
Accuracy	0.8341327072502654	0.8426966292134831	0.8431284588294481	0.8431256638157526	0.8480015652076694	0.8494913075074069
Precision	0.8303159213115249	0.8487601054053673	0.8450086284110809	0.8422079611526454	0.8487601054053673	0.8496308053813797
Recall	0.8333851884010003	0.8466089497052871	0.8429752796229250	0.8409212020118906	0.8472898214772309	0.8497149452253278
F1	0.8318477236985471	0.8476831628168261	0.8420616334853701	0.8409212020118906	0.8466089497052871	0.8482430103761518

corresponding to beta, lower gamma, and higher gamma bands and then flattened to get a vector of length 177 for each trial. The 25 principal components were extracted from the preprocessed data. To avoid over-congested figures we present plots for 5 first components only on Fig. 11. The complete plot of the 25 components is shown on Fig. 24 in Appendix. We can notice that despite the close proximity of samples from four emotion categories, the SVM model is capable of creating a very complex boundary in 25-dimensional space.

VIII. Additional Data Preprocessing to Improve the Accuracy of the Model

From Fig. 2 it is easy to notice that the spectral power density grows to infinity at the trial onset. It might be attributed to the event trigger influence. To reduce the effect of this outburst at the origin, we decided to cut off the first 10 time bins ($\sim$ 500 msec). It resulted in the increased accuracy for both the old and new models. The performance measures are summarized in Table I. Elimination of the spectral power density outburst at the stimulus onset leads to a slight increase in the accuracy of the old model. Furthermore, reducing the frequency range to beta, low gamma, and high gamma results in small improvements in the metrics. Notice that the new model performs slightly better than the old one and also follows a similar trend. In the case of the new model removing the spectral power density outburst at the stimulus onset (first 10 time bins ($\sim$ 500 msec)) and also removing the low frequencies leads to the increase of the accuracy to almost 85% (reaching the performance of the SVM model used in [3]). The average normalized confusion matrices for all three cases are shown on Fig 22 for old and on Fig 23 for new models correspondingly (in Appendix).

The removal of the spectral power density outburst at the stimulus onset also slightly improves the performance of the SVM model from [3]. While the reported accuracy of the original SVM model was about 0.85, the same model applied to the data with the first 10 time bins ($\sim$ 500 msec) removed achieves 0.86 accuracy after 10 folds cross-validation, thus resulting in almost 1% increase in accuracy. The other performance metrics also increase by around 1%.

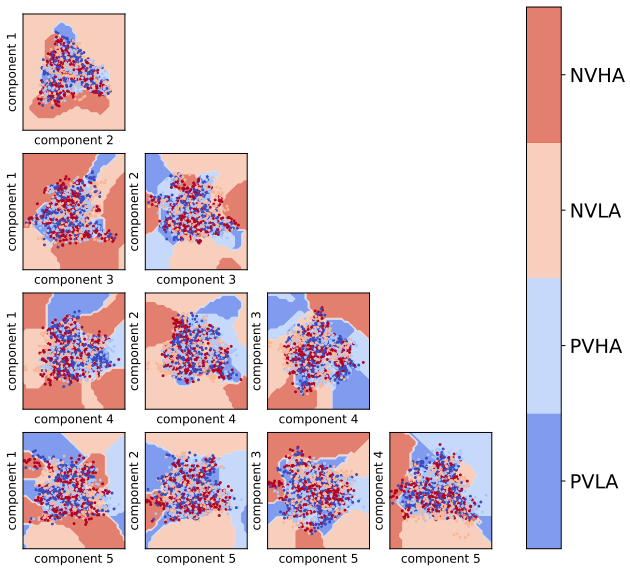


Fig. 11: Visualization of decision boundaries drawn by SVM with RBF ($\gamma = 0.1$) for first 5 principal components.

IX. Proposed Pipeline for Emotion Classification

Fig. 12 shows a proposed pipeline for emotion classification from EEG data with SVM and CNN models. Initially, the raw EEG data are passed through 0.1–85 Hz band-pass filter and 50 Hz notch filter and corrected with a common average reference method. Then the power spectral density of 1–80 Hz frequency bins was extracted using the Short-time Fourier Transform (STFT) with 500ms window. The processed EEG data represents power density from 2671 trials, each containing 160 time points recorded from 59 sensors and distributed across 80 frequency bins. Next, we check the data for artifacts caused by the stimulus onset and remove the lower frequencies as theta (4-8 Hz), slow alpha (8-10 Hz), and alpha (8-12 Hz). For CNN inputs, the 4D tensor of $68 \times 150 \times 59 \times 2671$ with artifacts removed is averaged across time and an additional singleton dimension is added to make it suitable for processing with CNN. For SVM the $67 \times 150 \times 59 \times 2671$ tensor is averaged across the time and three frequency bins, corresponding to beta, lower gamma, and higher gamma bands. Afterwards, the newly-obtained $3 \times 59 \times 2671$ tensor was flattened to get a vector of length 177 for each trial. Before supplying to the SVM model we

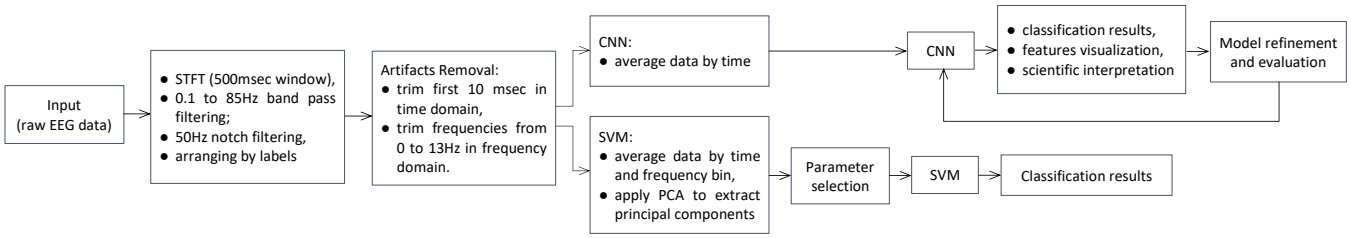


Fig. 12: Proposed pipeline for emotion classification from EEG data.

perform dimensionality reduction by applying PCA. Another approaches for dimensionality reduction as feature selection are also possible, however, experiments demonstrate that PCA allows to achieve higher accuracies. To improve the performance of the SVM model, the parameter selection with grid search is conducted. For CNN model visualization of hidden layers can provide information about the processes that are responsible for eliciting the human emotions and classifying them to the categories. This information can be used for further improvement of the model architecture and fine tuning the model parameters. In overall, the CNN model is more interpretable in the sense it allows to investigate the feature maps and activations of hidden layers, while it is hard to infer any specific patterns from the boundaries drawn by SVM in the multidimensional space.

X. Conclusion

In this project, we address the human emotion classification problem from the circumplex model perspective. We supply EEG data into the CNN model and investigate how the non-image data are processed in the hidden layers and how the activation maximizations of the output layer can be interpreted. While CNNs have been actively used for emotion recognition [16] [17], [20], to our best knowledge, there are no works that address the visualization of the hidden layers in the CNN model trained to recognize emotions. The understanding of how the CNN extracts features and makes the decision might shed some light on the understanding of the neurophysiological processes responsible for the elicitation of human emotions.

We start with visualizing the preprocessed EEG data available to us (bandpass and notch filter, and power spectral densities extracted). We present the time course of the data by grouping them into three frequency bands and averaging over channels and trials (Fig. 2). In our case, the duration of the trial (8 sec) is not enough to introduce significant changes in the elicited emotion, thus, the spectral power densities can be considered relatively stable. Moreover, apparently, emotion markers are sparsely distributed across the time, since experiments with classifying emotions for different timepoints lead to the accuracies of 0.5 and less, while using data averaged over time allows to reach classification accuracies around 0.85. It might also suggest the presence of the random noise distributed over time, which gets eliminated after averaging. For these reasons the Long-Short Term Memory networks also achieve low emotion recognition accuracy (two LSTM layers with 256 units each were able to achieve only 0.6 accuracy after 100 epochs).

The visualization of the preprocessed EEG data suggested another approach for preprocessing the data by removing the first 10 timebins (~ 500 msec) of each trial, when the power spectral densities are not stable. In fact, at the beginning of each trial, the outburst of the power spectral density is observed, which might introduce additional noise. Removal of the first 10 timebins increases the accuracy of the models by almost 1% which might be a significant improvement in the emotion recognition domain. Furthermore, focusing only on the beta, lower gamma, and higher gamma bands allows slightly increasing the accuracies, which shows that the theta, slow alpha, and alpha bands do not significantly contribute to the emotion classification. Interestingly, in [24] the authors found that emotion recognition is more associated with high-frequency oscillations in higher gamma band (51–100Hz) of EEG signals rather than low-frequency oscillations (0.3–49Hz). To test this observation we performed classification with ether beta (13Hz – 30Hz), or lower gamma (31Hz – 50Hz), or higher gamma (51Hz – 79Hz). The classification accuracies for beta, lower gamma, and higher gamma were 0.79, 0.82, and 0.84 correspondingly. Apparently, the higher frequencies allow to predict emotions more accurately, however, in our case the difference is not that significant, which can be explained by the different approaches and datasets used to collect the data. Moreover, as it has been concluded from the visualization of CNN layers, some emotional categories such as NVLA might be associated with oscillations in alpha (8-12 Hz) and beta (13Hz – 30Hz) bands, and focusing only on the high frequencies might result in a decrease in accuracy for some emotion classes.

The main focus of this project is to interpret the emotion classification in the CNN model and to offer an improved architecture. We visualized feature maps extracted by the old CNN model and noticing that the max-pooling layer decreases the "resolution" of the final feature map, we decided to remove it from the architecture. In the meantime, the feature maps of the convolution layers show that the filters learn patterns of the activations of sensors for different frequencies. Some filters consider activations of single sensors, others – groups of sensors. To capture the variety of activation patterns we decided to increase the number of filters (and it is also recommended for increasing the prediction accuracy of the model). The final convolution layer of the new model (Fig. 7, b.)) is capable of better discerning between the four emotion categories, which results in the increase of the accuracy of the model. Furthermore,

the activation maximization of the dense layer of the new model (Fig. 10), while showing similar patterns as of the old one (Fig. 9), provides a more clear idea about which frequency ranges are responsible for eliciting different types of emotions. Thus, we conclude from Fig. 10 that the NVLA is characterized by activations in the alpha and beta bands, while NVHA and PVHA can be described by the activations in the lower gamma and in high gamma bands. Notice that we cannot make similar conclusions about the activation of brain areas (under corresponding sensors) due to the poor spatial resolution of EEG.

We also proposed a pipeline for processing and analyzing the EEG data for creating interpretable models for human emotion recognition. The SVM model requires additional dimensionality reduction and does not produce a meaningful interpretation of the classification process. All we know is the SVM creates a very complicated boundary in 25-dimensional space. However, SVM is faster than CNN and can be used to validate the CNN's prediction. In contrast, CNN does not need any additional processing of the data (it can even work directly with raw EEG data, however, we were not able to test it, since we have access to the preprocessed data only). CNN also allows to create a meaningful interpretation of what is learned by the model and make some conclusions about the features associated with different emotion categories. Moreover, using CNNs will open an opportunity to integrate data from psychophysical signals that indirectly reflect the emotional state of the subject, such as heartbeat rate, skin conductance, and etc. This will allow to elicit emotions with stimuli of other modalities, for example, thermal or haptics.

For future work, it would be interesting to address the relationship between EEG oscillation frequencies and emotional category. To answer the question of which brain areas correspond to the elicitation of a particular emotion category, it is necessary to redesign the experimental setup to either collect more EEG data or to combine the EEG data with other recordings, that ensure high spatial resolution, e.g. fMRI scans. It will also require reconsidering the processing of EEG data for source recognition. Another direction to look into was inspired by our attempt to extract the meaningful patterns with t-SNE (Fig. 5 and Fig. 6). We can see that t-SNE produces separate agglomerations of small clusters corresponding to all four emotion categories. This implies that the current circumplex model does not capture the variety of human emotion categories. Each agglomeration might correspond to a level intensity of each of the four categories. It suggests that the current classification model can be transformed into a regression model where arousal and valence can be mapped into a quantitative number (positive and negative). The trained model can be used in several applications where the intensity of emotion is important, including health care, gaming and entertainment, and biofeedback.

REFERENCES

[1] J. Panksepp, "Affective neuroscience the foundations of human and animal emotions," New York, 2004.

[2] L. F. Barrett, B. Mesquita, K. N. Ochsner, and J. J. Gross, "The experience of emotion," *Annu. Rev. Psychol.*, vol. 58, pp. 373–403, 2007.

[3] V. Babushkin, W. Park, M. Hassan Jamil, H. Alsuradi, and M. Eid, "EEG-based classification of the intensity of emotional responses," in *10th International IEEE EMBS Conference on Neural Engineering*, 2021.

[4] A. Bhardwaj, A. Gupta, P. Jain, A. Rani, and J. Yadav, "Classification of human emotions from EEG signals using SVM and LDA classifiers," in *2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN)*, 2015, pp. 180–185.

[5] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[6] N. Dar, M. Akram, S. Khawaja, and A. Pujari, "CNN and LSTM-based emotion charting using physiological signals," *Sensors*, vol. 20, p. 4551, 08 2020.

[7] Y. Zhao, X. Cao, J. Lin, D. Yu, and X. Cao, "Multimodal emotion recognition model using physiological signals," 2019.

[8] S. Siddharth, T.-P. Jung, and T. J. Sejnowski, "Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing," 2019.

[9] A. T. Sohaib, S. Qureshi, J. Hagelbäck, O. Hilborn, and P. Jerčić, "Evaluating classifiers for emotion recognition using EEG," in *Foundations of Augmented Cognition*, D. D. Schmorrow and C. M. Fidopiastis, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 492–501.

[10] P. Rani, C. Liu, N. Sarkar, and E. Vanman, "An empirical study of machine learning techniques for affect recognition in human–robot interaction," *Pattern Analysis and Applications*, vol. 9, 05 2006.

[11] Y.-P. Lin, C.-H. Wang, T.-L. Wu, S.-K. Jeng, and J. Chen, "EEG-based emotion recognition in music listening: A comparison of schemes for multiclass support vector machine," 04 2009, pp. 489–492.

[12] R. Horlings, D. Datcu, and L. Rothkrantz, "Emotion recognition using brain activity," 01 2008, p. 6.

[13] V. Anh, M. Van, B. Ha Bang, and T. Huynh Quyet, "A real-time model based support vector machine for emotion recognition through EEG," 11 2012, pp. 191–196.

[14] P. Petrantonakis and L. Hadjileontiadis, "Emotion recognition from brain signals using hybrid adaptive filtering and higher order crossings analysis," *Affective Computing, IEEE Transactions on*, vol. 1, pp. 81–97, 07 2010.

[15] Y. Zhu, S. Wang, and Q. Ji, "Emotion recognition from users' EEG signals with the help of stimulus videos," vol. 2014, 07 2014, pp. 1–6.

[16] H. Mei and X. Xu, "EEG-based emotion classification using convolutional neural network," in *2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, 2017, pp. 130–135.

[17] R. Alhalaseh and S. Alasasfeh, "Machine-learning-based emotion recognition system using EEG signals," *Computers*, vol. 9, no. 4, 2020.

[18] Y.-H. Kwon, S.-B. Shin, and S. Kim, "Electroencephalography based fusion two-dimensional (2D)-convolution neural networks (CNN) model for emotion recognition system," *Sensors (Basel, Switzerland)*, vol. 18, 2018.

[19] S. Koelstra, C. Muhl, M. Soleymani, J. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis ;using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[20] J. Liu, G. Wu, Y. Luo, S. Qiu, S. Yang, W. Li, and Y. Bi, "EEG-based emotion classification using a deep neural network and sparse autoencoder," *Frontiers in Systems Neuroscience*, vol. 14, p. 43, 2020.

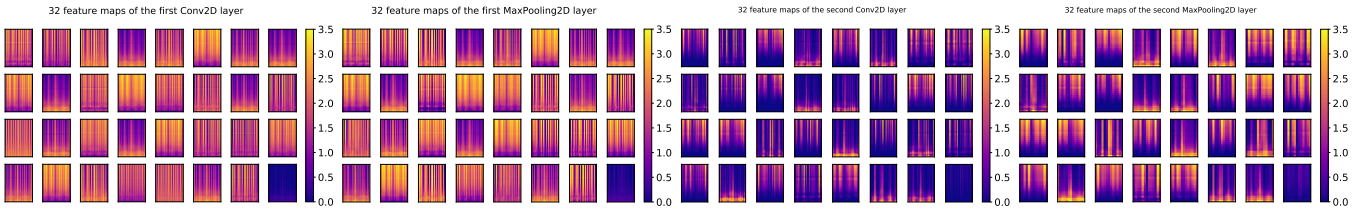
[21] N. Liu, Y. Fang, L. Li, L. Hou, F. Yang, and Y. Guo, "Multiple feature fusion for automatic emotion recognition using EEG signals," 04 2018, pp. 896–900.

[22] L. van der Maaten and G. Hinton, "Viualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 11 2008.

[23] A. Bilal, A. Jourabloo, M. Ye, X. Liu, and L. Ren, "Do convolutional neural networks learn class hierarchy?" *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 152–162, 2018.

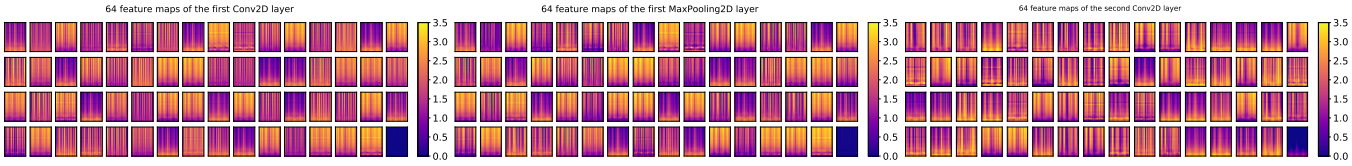
[24] Z. Gao, X. Cui, W. Wan, and Z. Gu, "Recognition of emotional states using multiscale information analysis of high frequency EEG oscillations," *Entropy*, vol. 21, no. 6, 2019. [Online]. Available: https://www.mdpi.com/1099-4300/21/6/609

XI. Appendix



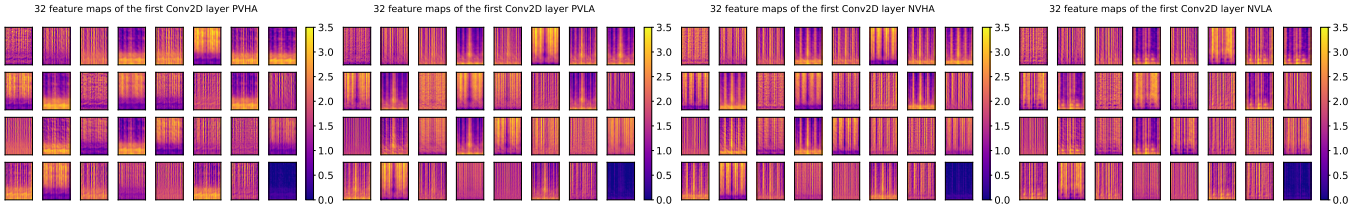
a.) b.) c.) d.)

Fig. 13: Feature maps averaged over all classes of the a.) first convolution, b.) first max pooling, c.) second convolution, d.) second max pooling layers of the model in [3] (Fig. 7, a.)).



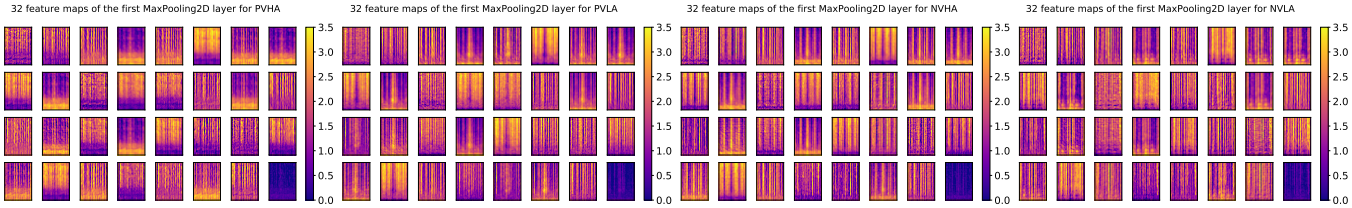
a.) b.) c.)

Fig. 14: Feature maps averaged over all classes of the a.) first convolution, b.) first max pooling, c.) second convolution layers of the proposed model (Fig. 7, b.)).



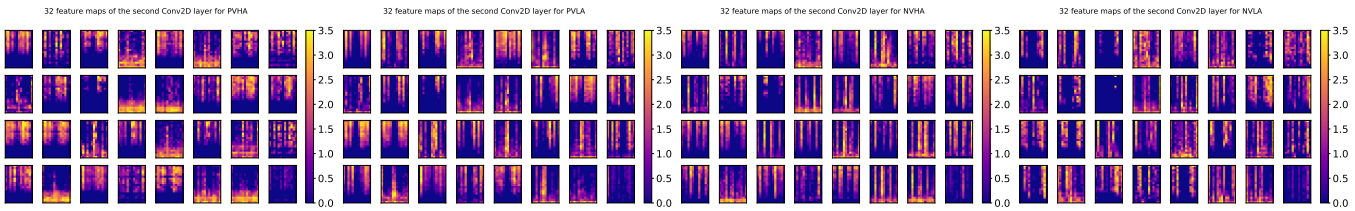
a.) PVHA b.) PVLA c.) NVHA d.) NVLA

Fig. 15: The features maps visualized for the first convolution layer of the old model for a single instance of a correctly predicted sample for each of four categories.



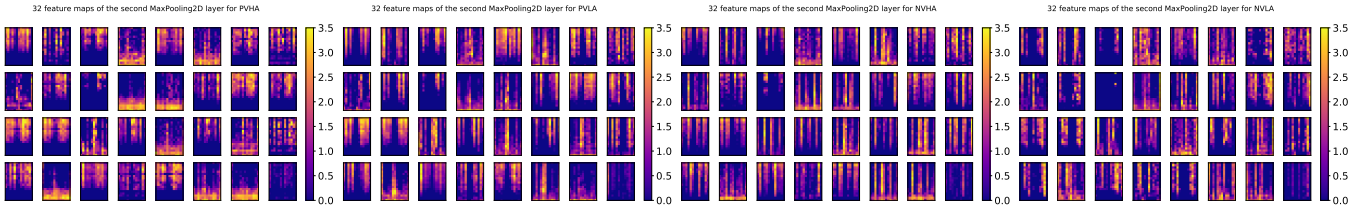
a.) PVHA b.) PVLA c.) NVHA d.) NVLA

Fig. 16: The features maps visualized for the first max pooling layer of the old model for a single instance of a correctly predicted sample for each of four categories.



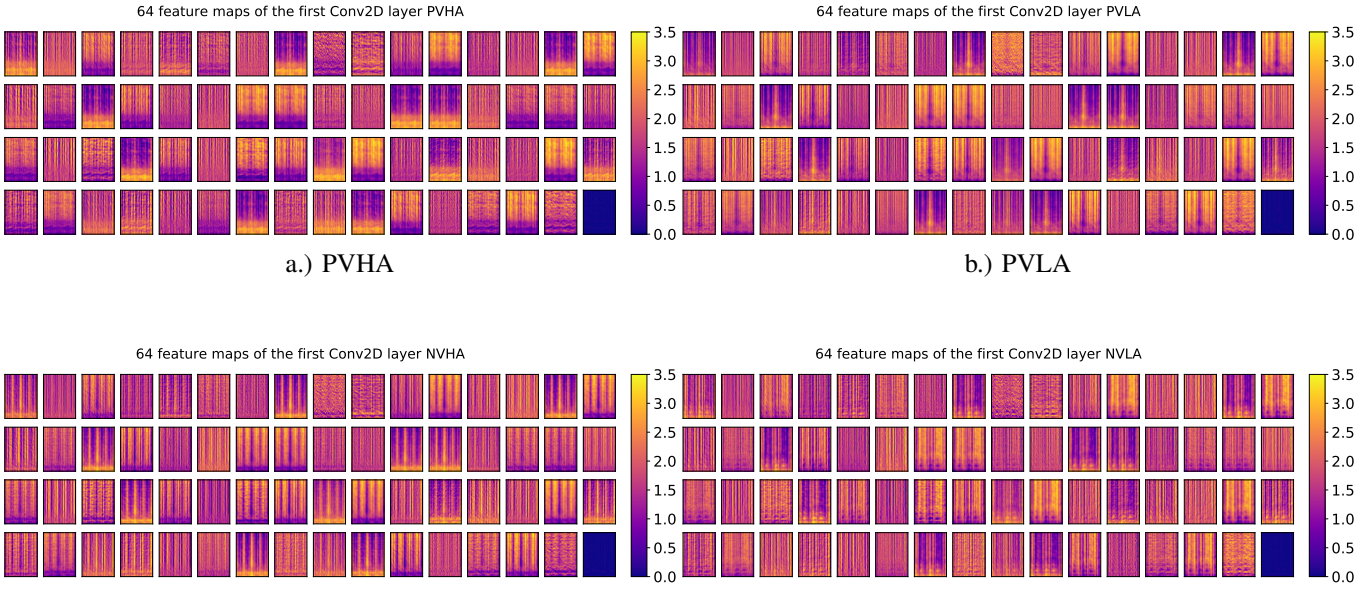
a.) PVHA b.) PVLA c.) NVHA d.) NVLA

Fig. 17: The features maps visualized for the second convolution layer of the old model for a single instance of a correctly predicted sample for each of four categories.



a.) PVHA b.) PVLA c.) NVHA d.) NVLA

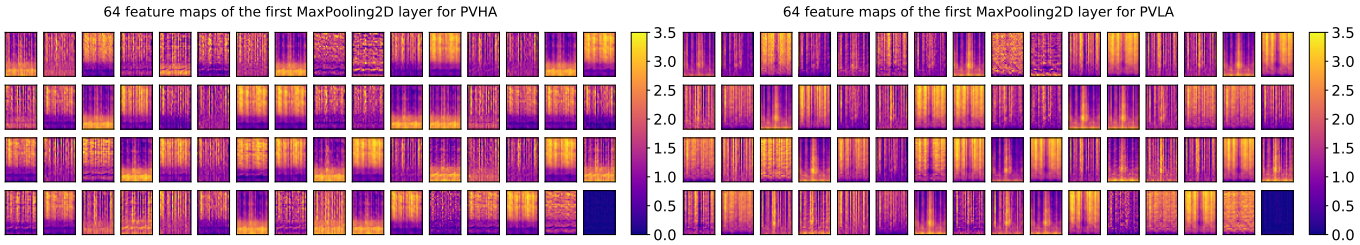
Fig. 18: The features maps visualized for the second max pooling layer of the old model for a single instance of a correctly predicted sample for each of four categories.



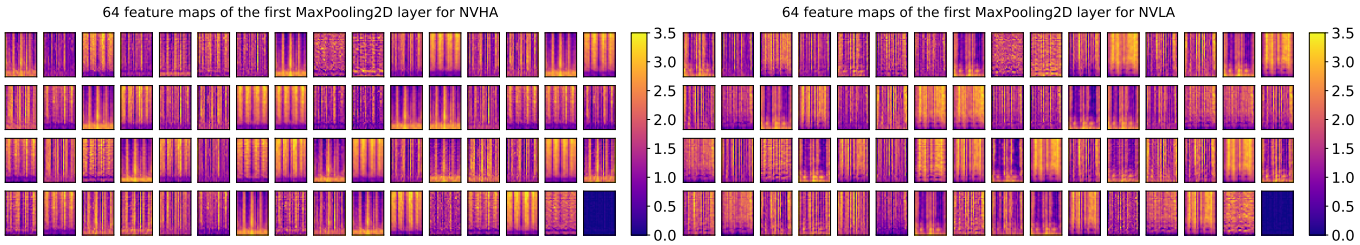
a.) PVHA b.) PVLA

c.) NVHA d.) NVLA

Fig. 19: The features maps visualized for the first convolution layer of the new model for a single instance of a correctly predicted sample for each of four categories.

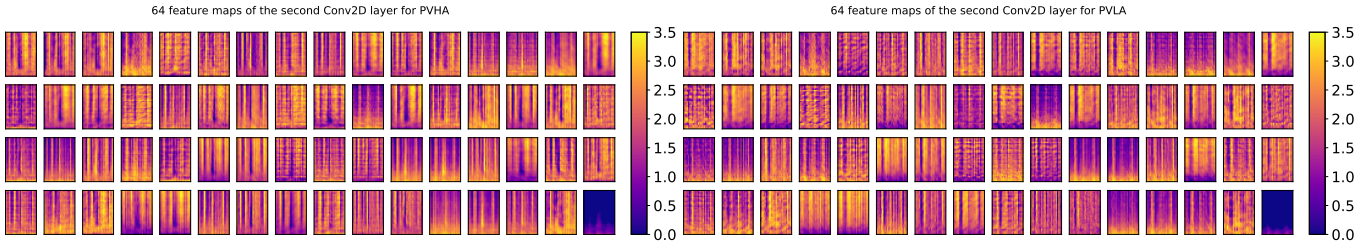


a.) PVHA b.) PVLA

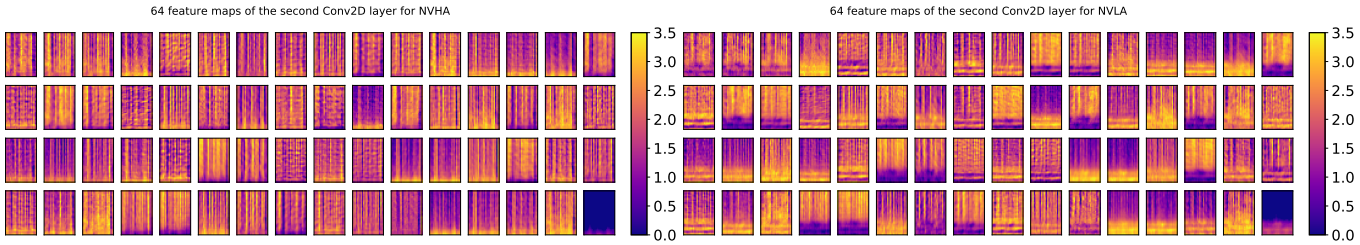


c.) NVHA d.) NVLA

Fig. 20: The features maps visualized for the first max pooling layer of the new model for a single instance of a correctly predicted sample for each of four categories.

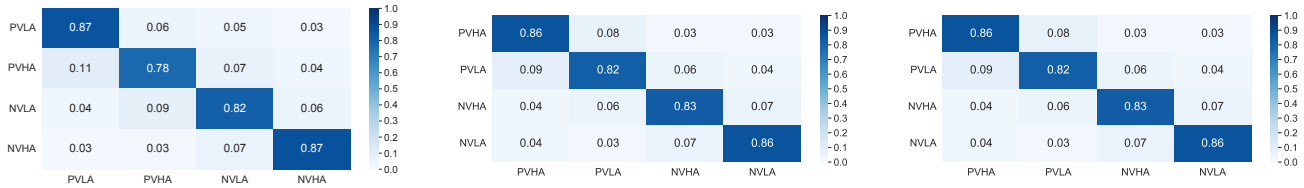


a.) PVHA b.) PVLA



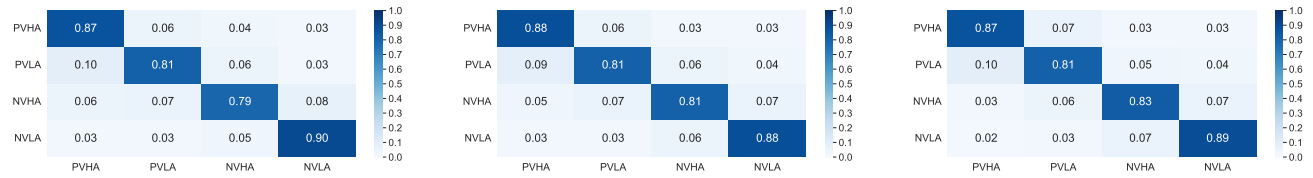
c.) NVHA d.) NVLA

Fig. 21: The features maps visualized for the second convolution layer of the new model for a single instance of a correctly predicted sample for each of four categories.



a.) b.) c.)

Fig. 22: Normalized confusion matrices averaged over 10 folds for old model evaluated on a.) full data, b.) first 10 timebins ($\sim$ 500 msec) trimmed, c.) first 10 timebins ($\sim$ 500 msec) and frequencies trimmed to beta, low and high gamma bands.



a.) b.) c.)

Fig. 23: Normalized confusion matrices averaged over 10 folds for the new model model evaluated on a.) full data, b.) first 10 timebins ($\sim$ 500 msec) trimmed, c.) first 10 timebins ($\sim$ 500 msec) and frequencies trimmed to beta, low and high gamma bands.

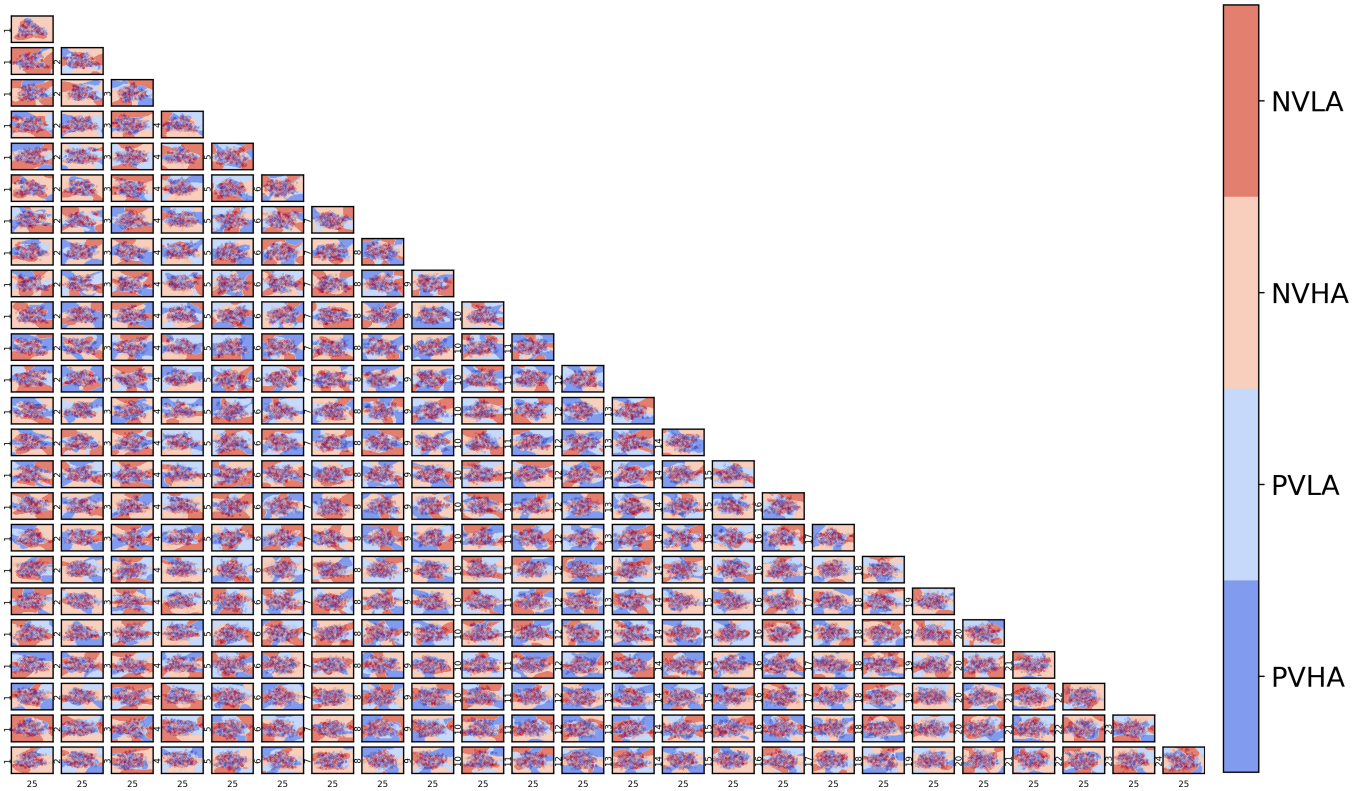


Fig. 24: Visualization of decision boundaries drawn by SVM with RBF ($\gamma = 0.1$) for 25 principal components.