Emotion Recognition Based on Multimodal Physiological Signals Through Deep Learning

Nelson Hidalgo
Department of Brain and Cognitive Science and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139
nelsonh@mit.edu

Abstract

The field of emotion recognition aims to automatically quantify human emotional states based on behavioural and physiological data. Interesting applications of this area of study include the early detection and relieve of strong negative emotions, and the diagnosis of neurological disorders that affect people's emotional well being. Much of the current emotion recognition research aims to differentiate from 3 emotional states, positive, negative, or neutral, and further work is needed to classify more nuanced emotional states. In addition, there is growing interest for classifying emotions based on wearable devices that are available beyond laboratory settings. My research develops a convolutional neural network algorithm that is able to estimates participants' ratings of 9 distinct emotions based on EEG, EDA, and BVP data collected with commercial-grade devices. My algorithm performs a regression estimate that is within 0.30 of participants' self-reported emotions on a 0 to 4 scale.

1 Introduction

Physiological computing is a field that aims to assess human physical and cognitive states based on biological signals or external behaviour. An important area of physiological computing is emotion recognition. Emotion recognition, also known as affect recognition, aims to better undrestand and classify human emotional states. Emotion classification is important in applications such as interventions to alleviate negative emotions (i.e. distress, sadness, etc.), affective computing, and diagnosis or analysis of neural disorders such as autism, depression, and anxiety (Zeng et al. [2019]).

In order to identify emotions, some of the most commonly used metrics starting with the most commonly used are the following: cardiovascular signals such as heart rate (HR) that can be estimated from plethysmography (PPG) or blood volume pulse (BVP), electrodermal activity (EDA) which relates to involuntary sweating reactions, video recordings of the person's face, and electrical activity in the brain obtained through electroencephalography (EEG) (Jemioło et al. [2022]). These metrics are often recorded altogether in order to better understand emotions while people watch videos that have been demonstrated to evoke specific emotions.

The growth of emotion recognition research was fueled by the availability of more accessible devices for the recording of the aforementioned bio-electrical signals. This led to the creation of open-source datasets that can be used by the machine learning and affective computing community to classify emotions (Saganowski et al. [2022], Jemioło et al. [2022]). Support vector machines have been the most commonly applied machine learning technique to classify emotions based on hand-crafted features. Deep learning has been applied in less than 12 percent of the papers in the field (Jemioło et al. [2022]).

Preprint. Under review.

The need for large amounts of data in order to train deep learning models is part of the reason why it is challenging to apply deep learning in this area of research, in which data is often scarce. Zeng et al. [2019] and Song et al. [2018] applied convolutional neural networks (CNNs) that acheived higher results than state of the art in differentiating between positive, negative, or neutral emotional states. A limitation of these studies is that they only differentiate between three emotional states. In addition, the datasets that they studied use equipment that is limited to laboratory settings, unavailable for every-day commercial use. Further work is needed to apply deep learning to the classification of more complex emotional states from commercially available devices.

Emognition, by Saganowski et al. [2022], is a dataset that addresses the need for using commercially available devices for recording HR, EDA, and EEG in emotion recognition research for real-world applications. This dataset also contains participant's self-reported emotions across categories including amusement, anger, awe, disgust, enthusiasm, fear, liking, sadness, and surprise. In order to classify such complex emotions, my research implements a CNN architecture which takes as input filtered physiological data, which is not modified with any hand-crafted features, and outputs a regression estimate of each emotional state.

My CNN architecture, available at https://github.com/nelsonalbertohj/EmotionNet, is inspired by previous work to classify EEG signals from visually evoked potentials, and it implements some key properties: (1) having a first convolutional filter with linear activations that serves to automatically extract spectral information from the signal (Khok et al. [2020], Waytowich et al. [2018]), and (2) using depth-wise and separable confvolutions that minimize the number of parameters in the network and improve training on small datasets (Khok et al. [2020]). My hypothesis is that my CNN architecture will be able to approximate well the emotional states reported by the participants. It will be able to make continuous regression estimates of all 9 emotional states reported in the study as well as discrete estimates of the most predominant emotional state.

2 Methods

2.1 Data Acquisition and Preparation

The Emognition dataset, by Saganowski et al. [2022], contained data recorded from 43 participants while they watched videos approximately 2 minutes in length. The dataset contains data from the Empatica E4, the Samsung Galaxy Watch, the Muse 2 headset, and the Samsung Galaxy s20 with video footage of the participants' reactions. The data from Muse 2 was used to extract raw EEG signals from channels AF7, AF8, AF9, and AF10, which are positioned in the forehead and through the top of the ears. The data from Empatica E4 was used to obtain HR, BVP, and EDA signals. These specific signals were chosen as they are the most commonly used in emotion research (Jemioło et al. [2022]). The response of the participants accross the 9 emotion are used as labels during the regression task. Given that this dataset is unbalanced because the videos for suprise and disgust are

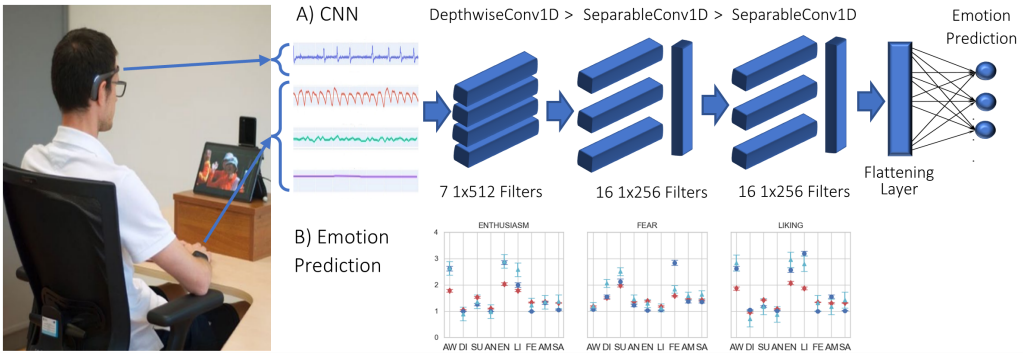

Figure 1: Diagram of the experiment and algorithm: (A) EEG data recorded from Muse headset and BVP, HR, and EDA data recorded from the Empatica E4 are passed through the 1D CNN architecture and a regression prediction across 9 emotions is produced, (B) the emotion prediction output is comapred with the participants' self-reported emotions.

only 1 minute long, while the others are about 2 minutes long, I duplicated samples from the shorter recordings as needed so that all labels were equally represented during training and testing.

The architecture shown in 1.A was used to analyze the physiological signals. The input to the CNN consisted of 7 channels. The first 4 were EEG channels filtered using a 4^{th} order Butterworth filter to values between 5 and 55 Hz, a frequency that contains relevant EEG frequency bands: some of Theta (4-8 Hz) associated with relaxation and meditation, Alpha (8-12 Hz) associated with creativity and relaxation, Beta (16-32 Hz) associated with concentration. The fifth channel was blood volume pulse, BVP, filtered using biosppy BVP processing tool by Carreiras et al. [2015–] which uses a Butterworth bandpass filter between 1 and 8 Hz.

The next channel was heart rate, HR, estimated from the BVP based on a beat count from the BVP signal (Carreiras et al. [2015–]). The final channel contained EDA signal filtered to frequencies bellow 5 Hz with a low-pass Butterworth filter and smoothed by a 192 samples moving average window (Carreiras et al. [2015–]). In order to keep the input signal size consistent, the EEG signal was down-sampled from 256 Hz to 64 Hz, the BVP signal was kept at 64 Hz, and the EDA signal was up-sampled from 4 Hz to 64 Hz. As a way to augment the data and capture enough samples to identify emotions, the data from each 2 second trials was divided into 30 second windows with 25 seconds of overlap, resulting in 24 samples per video session.

2.2 CNN model

The first layer of the CNN model is a convolutional layer with a filter of dimension 512. The first convolutional layer uses linear activations to perform a type of band-pass filtering according to the research of Waytowich et al. [2018] and Khok et al. [2020]. This first layer is a depth-wise convolution that acts on each channel independently, meaning that there is no information mixing across channels at this point in the network.

The following layers are separable convolutions, consisting in one layer that convolves across the time axis and a subsequent one that convolves across the channels axis. Afterwards, batch normalization is applied before passing the pre-activation outputs through exponential linear unit (ELU) activations, Equation (1). ELU activations have resulted in performance improvements in neural networks for EEG classification (Lawhern et al. [2018]). The implementation of depth-wise separable convolutions reduces the number of weights that need to be trained when compared to traditional 2D convolutions. This property is desirable for this kind of application where small amounts of data are available.

$$ELU(x) = \begin{cases} x & x > 0 \\ \alpha(e^x - 1) & x <= 0 \end{cases} \tag{1}$$

After flattening the outputs from the convolution layers, the network has a fully-connected layer with 9 output units with linear activations. These output units perform a regression estimate for each of the 9 emotional states that were reported in the Emognition dataset on a scale from 0 to 4. Gradient descent with the Adam optimizer (Kingma and Ba [2014]) is used to train the weights of the network based on the mean squared error, Equation (2), loss function. The target labels are the discrete ratings that the participants reported. On a separate part of the experiment, my CNN was trained to output discrete, rather than regression, predictions of the most predominant emotional state. In this case, softmax output activation was used and categorical cross-entropy was the loss function of choice.

$$MSE(\hat{y}) = \frac{1}{9}\sum_{i=1}^{9}(\hat{y_i} - y_i)^2 \tag{2}$$

The hyper-parameters of the network were determined by changing the filter sizes from 64, 128, 256, 512, to 768 for the first filter and by adding separable convolutional layers. The chosen filter size was 512, which may be more appropriate in this application where large portions of the signal, 30 seconds or 1920 samples, are used per data point. The physiological signal processing literature has demonstrated performance gains in using larger convolutional filters (Khok et al. [2020]).

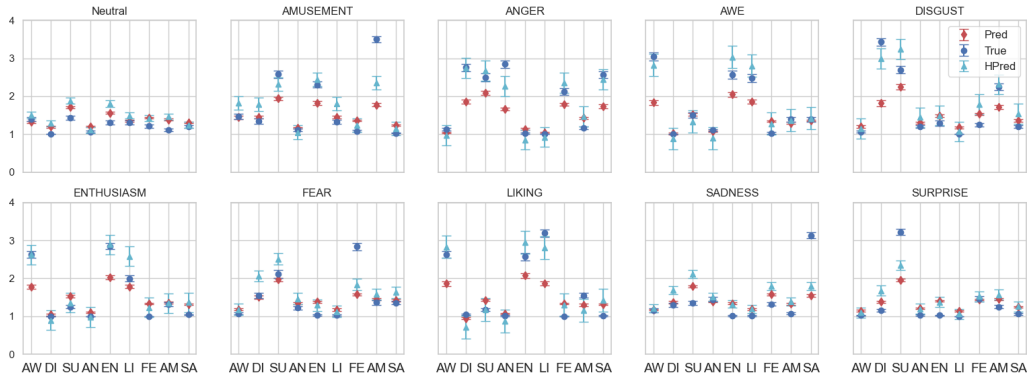

Figure 2: The model estimations of awe (AW), disgust (DI), surprise (SU), anger (AN), enthusiasm (EN), liking (LI), fear (FE), amusement (AM), and sadness (SA) across different target emotion on a scale from 1 to 4 are shown. The blue markers represent the averaged self-reported emotions across all participants, the red markers are the averaged model estimation of participants emotions across all participants, and the light blue markers are the predictions made by the model in the top 25 % of data points with most variance in regression estimates.

3 Results

3.1 Regression Task

The results in Figure 2 show the average regression predictions versus the average of participants' self-rated emotions for each target emotion. The target emotion refers to the emotion that each video intended to evoke in the participant. All emotions were rated as: 0-not at all, 1-slightly, 2-moderately, 3-very, or 4-extremely. For example, the Neutral figure corresponds to the moments where participants were watching the 2 seconds-long Neutral stimulus: the blue markers were the average self-reported emotion from all participants after watching the video, the red markers are the average network predictions for every 30 seconds fragment of physiological data extracted during the 2 seconds-long stimulus, and the light blue markers are similar to the red markers except that they are chosen from predictions that showed most variance.

The markers from predictions with most variance are reported to take into account that the participant may not experience the same emotional state throughout the video. Thus, the results with most variance are expected to highlight predictions during moments of most intense emotions. A more detailed perspective on exact differences between predicted emotions and self-reported emotions is presented in Table 1 for the difference between overall predictions (the red marker) and the true labels, and in Table 2 for the difference between predictions with most variance (the light blue marker) and the true labels. The target emotions are along the rows and the self-reported emotions along the columns.

Table 1: Difference Between CNN Prediction and Self-reported Emotions

Emotions	AW	DI	SU	AN	EN	LI	FE	AM	SA	Target Dif.	MAE
Neutral	-0.07	0.20	0.28	0.13	0.25	0.03	0.21	0.27	0.10	0.00	0.17
AMUSEMENT	-0.02	0.11	-0.64	0.05	-0.48	0.13	0.28	-1.73	0.21	-1.73	0.41
ANGER	-0.10	-0.89	-0.41	-1.18	0.11	0.05	-0.33	0.26	-0.83	-1.18	0.46
AWE	-1.21	0.02	0.02	-0.03	-0.52	-0.63	0.31	-0.09	-0.03	-1.21	0.32
DISGUST	0.14	-1.62	-0.45	0.10	0.19	0.18	0.29	-0.53	0.15	-1.62	0.40
ENTHUSIASM	-0.85	0.04	0.27	0.10	-0.83	-0.21	0.33	0.04	0.27	-0.83	0.33
FEAR	0.11	-0.03	-0.15	0.12	0.35	0.14	-1.26	0.09	0.06	-1.26	0.26
LIKING	-0.76	-0.10	0.26	0.06	-0.50	-1.32	0.34	-0.25	0.30	-1.32	0.43
SADNESS	0.01	0.07	0.43	-0.06	0.33	0.17	0.26	0.25	-1.59	-1.59	0.35
SURPRISE	0.12	0.23	-1.27	0.18	0.37	0.14	-0.02	0.20	0.18	-1.27	0.30
Average	-0.26	-0.20	-0.16	-0.05	-0.07	-0.13	0.04	-0.15	-0.12	-1.20	0.34

Table 2: Difference Between CNN Prediction with Most Variance and Self-reported Emotions

Emotions	AW	DI	SU	AN	EN	LI	FE	AM	SA	Target Dif.	MAE
Neutral	0.11	0.28	0.45	0.04	0.49	0.18	0.21	0.34	0.03	0.00	0.24
AMUSEMENT	0.36	0.46	-0.26	-0.06	0.13	0.49	0.15	-1.15	0.12	-1.15	0.35
ANGER	-0.16	0.00	0.17	-0.58	-0.17	-0.07	0.23	0.30	-0.12	-0.58	0.20
AWE	-0.22	-0.12	-0.17	-0.21	0.47	0.32	0.25	-0.02	0.03	-0.22	0.20
DISGUST	0.09	-0.44	0.55	0.24	0.20	0.07	0.54	0.07	0.33	-0.44	0.28
ENTHUSIASM	0.00	-0.11	0.10	-0.02	0.04	0.58	0.23	0.03	0.32	0.04	0.16
FEAR	0.11	0.52	0.39	0.24	0.26	0.09	-1.01	0.22	0.27	-1.01	0.35
LIKING	0.21	-0.34	0.01	-0.14	0.38	-0.38	0.31	-0.39	0.41	-0.38	0.29
SADNESS	0.05	0.39	0.76	0.06	0.31	0.18	0.47	0.31	-1.33	-1.33	0.43
SURPRISE	0.07	0.53	-0.87	0.17	0.35	0.06	0.06	0.32	0.20	-0.87	0.29
Average	0.06	0.12	0.11	-0.03	0.25	0.15	0.14	0.00	0.02	-0.60	0.28

3.2 Classification Task

My CNN was also tested on a classification task where it was trained to predict discrete the target emotion. This task is the same as identifying what video the participant was watching. Thus, rather than the output being a continuous value between 0 and 4, the output is one concrete target emotion, either Neutral, Amusement, Anger, Awe, Disgust, Enthusiasm, Fear, Liking, Sadness, and Surprise. The results of the classification task are shown in the confusion matrix in Figure 3. The overall classification accuracy of the CNN was 0.60. The accuracy for each condition was respectively: Neutral 0.32, Amusement 0.65, Anger 0.61, Awe 0.70, Disgust 0.53, Enthusiasm 0.59, Fear 0.56, Sadness 0.68, and Surprise 0.75. The confusion matrix indicates that the emotion that was most commonly mistaken was neutral, which may be due to the fact that emotions could be neutral for the most part in the course of a 2 minute video.

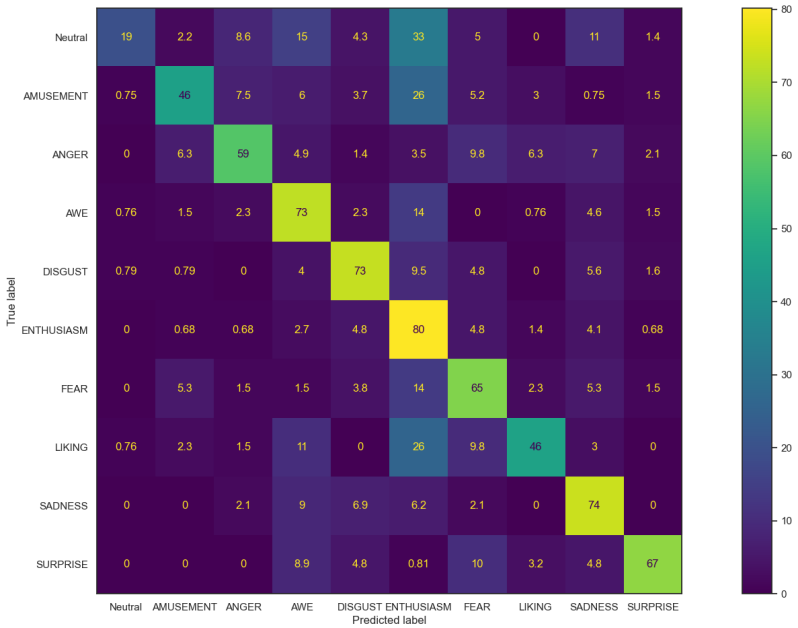

Figure 3: The confusion matrix shows along the central diagonal true positive classification percent. The values off the diagonal indicate the percent of wrongly classified emotions

4 Discussion

The results from the regression task present strong evidence that the model approximated well the emotional states that participants reported, with the caveat that there is a bias towards neutral emotions (ratings of 1-slightly or 2-moderately across all emotional states). This bias could result from the fact that emotional experience even while watching non-neutral video stimuli may be predominantly

neutral. The results from all emotion prediction differences averaged together, the last column of Table 1, show a mean absolute error of 0.34. This error rate is adequate given that emotion levels may be rounded to the correct discrete levels as long as the CNN estimate is within less than 0.5 of the true rating.

However, the bias towards neutral emotion becomes evident when looking at the average absolute difference between the self-reported emotion and the emotion predicted, which is shown in the second-to-last column of Table 1; this average difference of -1.20 indicates that the target emotion (the emotion that the video was intended to evoke) was under-estimated to be at least one entire rating point bellow what the participants reported. The consequence of this is that most of the data fragments from an emotion such as Amusement would be estimated as less than what the participant reported feeling.

In order to better understand this phenomenon, I also analyzed CNN predictions with top 25 % most variance, meaning highest difference between all predicted emotions, across each target emotion. These results are presented in Table 2 and Figure 2. I observe from the last column of Table 2 that the average absolute difference between the self-reported emotion and the emotion predicted is lower than for the aggregated predictions, which indicates a better fit to the overall reported emotional state. Most importantly, the average target difference reported in the second-to-last column of Table 2 dropped from -1.20 to -0.60. This result may indicate that in moments where the participant feels emotions most strongly while watching a video, the CNN model is able to differentiate between the emotional states.

My proposal about the relationship between moments of strong emotional differences and accurate model predictions needs further investigation that studies the signal exclusively around moments when the participants experience strong emotions. While Emognition does provide information about when the users made a facial expression when watching a video in their dataset, facial expressions are different across participants. Some people may experience emotions strongly but without making facial expressions. This may be addressed in future emotion datasets by having the participant annotate parts of the video from when they felt emotions most strongly.

In addition, my research explored how my CNN would perform if the task was framed as a discrete classification task intended to classify unique emotions. My results of 0.60 classification accuracy are significantly higher than chance level, which is 0.1 in this case. To my knowledge, there have not been classification models yet designed for this dataset to which I could compare my results. As shown in Figure 3, now that the network predictions are restricted to be one of the 10 target emotions and there are more non-neutral class labels than neutral class labels, the CNN is biased against predicting Neutral. In fact the precision for the Neutral label is 0.805, meaning that 80.5 % of all classes that are truly Neutral are classified as something else. This also indicates that there are fragments of the data during non-neutral stimuli which actually corresponds to a neutral state. As mentioned before, this needs to be investigated further in future work.

Further validation of the predictive power of my model could lead to interesting applications. For example, people could gain insights into their emotional state based on the data that is gathered with wearable devices, such as smart watches, while they were watching a video. On the other hand, testing of my model for clinical applications may help reveal when a hospitalized patient is feeling negative emotions, in which case different interventions may be design to uplift their mood and improve their experience.

5 Conclusion

In this investigation, I propose a CNN model that is designed to analyze physiological signals including brain signals, electrodermal activity, blood volume pulse, and heart rate in order to predict emotions. My regression model showed good estimations of emotional states, especially when considering those empotional states that were most salient from each other. In addition, the classification model built using a similar CNN architecture showed higher than chance classification across most emotional states with the exception of the neutral state. My results indicate that signals corresponding to neutral emotions are present in segments of recordings that are intended to be non-neutral. This phenomenon may be expected since participants will not necessarily experience the same level of emtions throughout a 2-minutes video. However, this highlights the need for more precise annotations of when the participants experienced strongest emotions. My future work aims to take this into

account by training and testing the model on data from moments when the participants are expected to have experienced greater differences in emotions.

References

Hong Zeng, Zhenhua Wu, Jiaming Zhang, Chen Yang, Hua Zhang, Guojun Dai, and Wanzeng Kong. Eeg emotion classification using an improved sincnet-based deep learning model. *Brain sciences*, 9(11):326, 2019.

Paweł Jemioło, Dawid Storman, Maria Mamica, Mateusz Szymkowski, Wioletta Żabicka, Magdalena Wojtaszek-Główka, and Antoni Ligęza. Datasets for automated affect and emotion recognition from cardiovascular signals using artificial intelligence—a systematic review. *Sensors*, 22(7):2538, 2022.

Stanisław Saganowski, Joanna Komoszyńska, Maciej Behnke, Bartosz Perz, Dominika Kunc, Bartłomiej Klich, Łukasz D Kaczmarek, and Przemysław Kazienko. Emognition dataset: emotion recognition with self-reports, facial expressions, and physiology using wearables. *Scientific data*, 9 (1):1–11, 2022.

Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3):532–541, 2018.

Hong Jing Khok, Victor Teck Chang Koh, and Cuntai Guan. Deep multi-task learning for ssvep detection and visual response mapping. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1280–1285. IEEE, 2020.

Nicholas Waytowich, Vernon J Lawhern, Javier O Garcia, Jennifer Cummings, Josef Faller, Paul Sajda, and Jean M Vettel. Compact convolutional neural networks for classification of asynchronous steady-state visual evoked potentials. *Journal of neural engineering*, 15(6):066031, 2018.

Carlos Carreiras, Ana Priscila Alves, André Lourenço, Filipe Canento, Hugo Silva, Ana Fred, et al. BioSPPy: Biosignal processing in Python, 2015–. URL `https://github.com/PIA-Group/BioSPPy/`. [Online; accessed <today>].

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.