EEG emotion recognition from facial expression stimuli with the presence of noisy labels

Konstantinos Azas

MSc in Data Science
The University of Bath
2021/2022

EEG emotion recognition from facial expression stimuli with the presence of noisy labels

Submitted by: Konstantinos Azas

Copyright

Declaration

This dissertation is submitted to the University of Bath in accordance with the requirements of the degree of Master of Science in the Department of Computer Science. No portion of the work in this dissertation has been submitted in support of an application for any other degree or qualification of this or any other university or institution of learning. Except where specifically acknowledged, it is the work of the author.

Abstract

This report portrays an investigation regarding EEG emotion recognition under facial expression stimuli, with main assumption the existence of 'noisy' labels. Data that were used, were collected with an EMOTIV EPOC headset that is utilised to gather EEG activity. Participants were given a sequence of images of people making a facial expression and simultaneously, they were writing the scores of emotions under their immediate reaction in terms of valence, arousal and dominance. For the signal processing pipelines, Independent Component Analysis (ICA) was used as a baseline approach and a deep generative model such as a Variational Autoencoder (VAE) was also implemented for determining the latent features within the EEG structure. Overall, autoencoders have achieved great results hence its implementation for EEG decoding is very promising. The datasets that the models were tested on, where our dataset and one public dataset (DEAP). Due to the existence of noisy labels, the two models were undertaken in a semi-supervised learning approach and a method known as 'pseudo' labelling was utilised to predict and replace the noisy labels with 'real' ones. Then, these labels are augmented in the original dataset with the rest that were already labelled and further regression modelling was implemented. In regression modelling, results were very good under the implementation of both ICA and VAE. Although, results were very similar which lacks the ability of a VAE to slightly outperform ICA. Additionally, the pseudo classifier worked impressively well but again pseudo labels weren't more accurate with the implementation of a VAE.

Contents

List of Figures

List of Tables

Acknowledgements

This dissertation topic was investigated by myself only. Although the many difficulties I have encountered throughout the process and that such a domain was completely new for me, I want to thank my supervisors Dr James Laird and Scott Wellington for the drive and support they gave me through this stressful and interesting period of the dissertation. Additionally, I want to thank every participant that help me do the data collection I was required to do for my investigation regarding how time consuming it was and for their willingness to undergo such a tiring task. I want to thank Dr Neal Hinvest from the department of Psychology of the University of Bath in spending some of his time with me in guiding me with my experimental paradigm for my data collection as well as understanding deeper certain domains of neuroscience that I wasn't really able to understand myself.

Chapter 1

Introduction

Emotions are something so common in an individual's everyday life. Regarding life events, there are occasions of incidents that people feel happy or sad. Not necessarily only with those but these two are the most general cases to describe a person's feelings. But is it possible to train a device to recognise this in a person? And to be more precise, is there eligibility to somehow quantify these feelings? For example, are we able to distinguish between common anxiety in a person and high levels of anxiety implying that he/she might suffer from anxiety disorder? So, instead of just identifying if an emotion exists by classification, regression methods must be performed where emotion can be quantified and estimated in order to model every possible emotion that exists.

Researchers mainly choose to investigate the detection of certain emotions from brain activity, using two separate approaches. The first is known as 'discrete' where an example of this, could be the classification of emotions like happiness, sorrow, excitement etc. But regarding emotions, there are so many factors that may affect them such as facial and bodily expressions or audio stimuli. Hence, this problem cannot be represented with single labels of emotions due to its complexity. For this, we may induce several 3 separate scales. The first one is known as valence and it distinguishes between positive/negative emotions, for example happiness/sadness. The second is known as arousal describes the intensity of the emotion, low arousal being boredom and high being excitement. The third and most important one is known as dominance and it refers to the degree of control of a certain emotion. For example, admiration is of high dominance because the person has control over that particular emotion. Indeed, fear is considered low dominance, because the person can't actually control when to fear and when not. It is an example of an instantaneous incident which also depends on its severity.

The reason why dominance is also included is because there is some ambiguity if valence/arousal are only used. An example of this issue could be with 'anger' and 'fear' which they are two different emotions but both have high level of arousal and negative valence [23]. A representation can also be shown in figure 1.1. This is the reason the emotions are represented in a dimensional space known as the PAD space where P stands for pleasure (valence), A stands for arousal and D stands for dominance [11]. A representation of the PAD space is shown in figure 1.2. Instead of classification, it would be more robust to use regression techniques to predict affective states on the PAD space. This approach is formulated as a 'dimensional' problem. And with the plots it makes better sense now on how to predict emotions using a regression models.

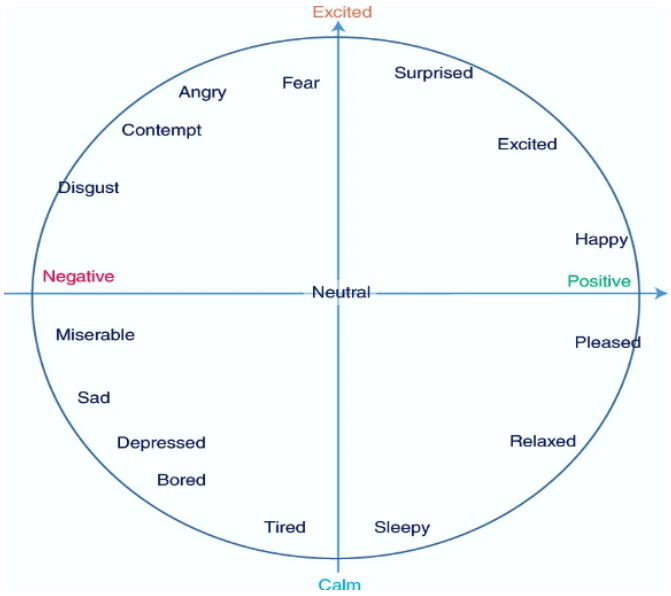

Figure 1.1: Valence-Arousal Scale from [40]. It categorises different types of emotions based only with valence and arousal.

To achieve all of the above, we have conducted an experiment where participants were selected for the gathering of their EEG signals using the Emotiv Epoc headset. Participants are being shown a database of images, showing people making a facial expression, basically describing an emotional state. Right after each picture, a few seconds were given for participants to complete a self-assessment manikin (SAM) [42] giving scores of valence-arousal-dominance to each picture they face. Then, an analysis will be performed by comparing the results from the SAMs with the Affective Norms for English Words (ANEW) [4]. The ANEW provides a set of normative emotional ratings for a large amount of words from the English dictionary. The purpose of using the ANEW in research is to compare results within our domain of investigation, in case of performing our own data acquisition. For each word, a rating of valence-arousal-dominance is being given. The scales for each dimension are given on a 9-point scale. These scores are obtained by some psychology students, where they were given a variation of the SAM, called the ScanSAM. The reason this analysis is performed is because there are a lot of inaccuracies for what a participant can write as scores. This leads for the data to be partitioned in two domains of 'labelled' and 'unlabelled' data separately. Then semi-supervised learning techniques will be performed to generate labels for the 'unlabelled' domain. This will occur simultaneously while we extract features from the EEG signals to reduce their dimensionality and determine the latent features that we look for within it, that are suitable for predicting emotional states. After this, a regression model will be trained to learn and predict emotional states in terms of valence-arousal-dominance. A variety of models will be implemented, such as Support Vector Machines for Regression (SVRs), Random Forests and k-Nearest Neighbours (k-NNs).

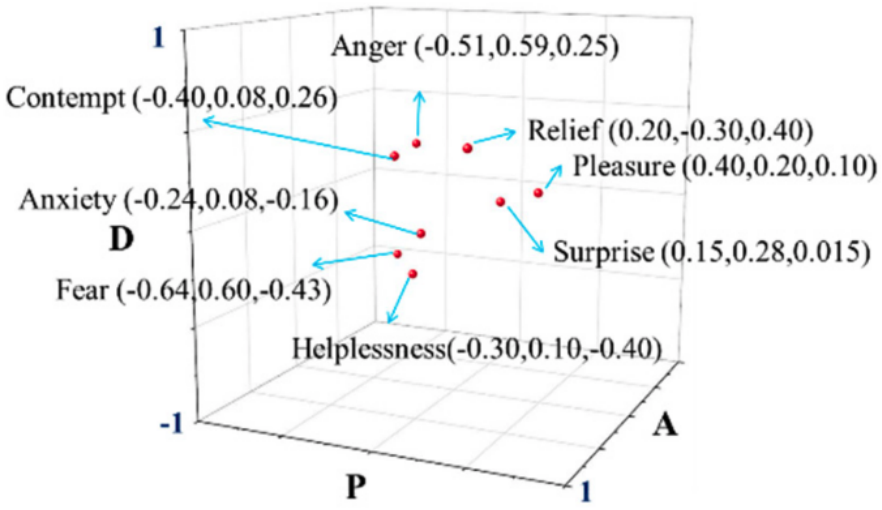

Figure 1.2: PAD space from [24], showing emotions in a dimensional model and their visualisation as vectors, with the inclusion of dominance.

Chapter 2

Background

2.1 Emotion Recognition

Emotion recognition represents the ability to encode an ensemble of sensory stimuli providing information about the emotional state of another individual [8]. Although, in terms of understanding people's feelings plays a crucial role in social interaction. Once a human sees people's reactions, he/she is capable of perceiving, anticipating and responding to them [18]. For example, if we observe someone our brain can gather a lot of information about that person's emotional state without the need of having prior knowledge to that individual. Mainly this information is perceived of what type of facial expression the individual does towards other, such smiling which depicts happiness, sharp eyebrows which depicts anger or wide open eyes and mouth which is an indication of fear. To generalise the term 'emotion' it can be represented in several dimensions such as valence, arousal and dominance as mentioned in chapter 1. Although, there are some others such as familiarity and liking which these mainly depend on the type of stimuli people interfere with.

2.2 Brain Computer Interfaces

Brain computer interfaces (BCIs) that acquires brain signals, analyzes them, and translates them into commands that are relayed to an output device to carry out a desired action [33]. Although, it is not capable of replacing explicitly systems within the human brain structure. BCIs can be considered as a sequential structure that attempts to complete several tasks in a manner. They allow a user to do actions using brain activity instead of muscle activity. More specifically, electroencephalography (EEG) is generated by the user and the BCI tries to decode the signal and retrieve an output that achieves the user's intention or action.

2.3 Eliciting emotions

There are many techniques that emotions can be elicited by humans. The most common technique is while using music. This will ensure that participants' emotions are going to be triggered and this is due to the release of neurotransmitters such as 'dopamine', 'noradrenaline' or even 'serotonin'. And to trigger those, it is done through the features from music known as 'Myriad features' that include harmony, timbre, interpretation or lyrics [16]. Instead, for this

investigation the stimuli we are going to use is pictures of people making a facial expression. Although there are some advantages, when a participant experience facial stimuli with respect to music listening. Facial emotions is considered one of the oldest communication skills. For example, female faces are utilized more efficiently in emotion recognition due to the fact that there was communication with their newborns during early development for example. Music is being processed different by humans. Triggering of emotions through music is achieved through some more general cognitive skills like attention, memory or even motivation [10]. Using facial stimuli, there are some other factors that might boost performance for emotion recognition tasks. Different races, through different facial stimuli is a key feature. For example, recognizing emotions in humans might be more efficient if they are being shown pictures of faces of their own race with respect to ones used from others [31].

2.4 Related Work

This section will consist of any work from past researchers as well as on what stimuli they have used to collect their data. As mentioned in Chapter 1, we are proposing to use facial stimuli in our experiments and also it is explained why it can be more suitable than using music. Although, there are various approaches on how researchers conducted experiments even with the usage of music. Of course, it depends on the researcher, on what emotions he/she is trying to elicit through certain stimuli.

Hadjidimitriou *et al.* [12], proposed a rare occasion where he used a joyful, pleasant stimulus as a consonant and the same stimulus with some manipulated counterparts as the dissonant. Stimuli duration were approximately 1 minute long. They also gave participants the opportunity to rate their likings for each stimulus. So their goal was to test whether an emotion can be detected if a person experiences a music clip that is regardless of familiarity, if it is familiar and if it is unfamiliar. For the classification task, he proposed mainly SVMs and k-NNs. And they have achieved classification accuracies of 87% for unfamiliar music, 91% for familiar music and 85% of regardless.

Thammasan *et al.* [38] decided to provide his participants with a music library, giving them the privilege to select songs on their own and before exposure they rated each song on how familiar they are with it. The investigation was based on the effects of familiarity/unfamiliarity based on EEG signals and it was implemented using their experimental dataset and the DEAP dataset. For investigating familiarity in music, the power spectral density (PSD) features were extracted using fast Fourier transform (FFT) and fractal dimension (FD) features using the Higuchi's algorithm. The machine learning models that were used to perform classification were SVMs, Multi-layer Perceptron (MLP) and the C4.5 model. The best results were obtained by extracting the FD features and implementing SVM on unfamiliar music with 87.80% classification accuracy.

Han *et al.* [13], decided to choose 165 western pop songs where he created 11 categories each of 15 songs, where each category corresponded to an emotion. They proposed to use multi-class SVR (11 classes) and achieved a classification accuracy of 94.55% and 92.73% for a Gaussian mixture model (GMM). This was done by extracting 7 distinct features from music mapped into 11 emotion categories on Thayer's emotion model. Then SVR and GMMs were used to predict valence and arousal values.

But for the sake of utilizing facial stimuli instead, we are going to implement the experimen-

tal paradigm proposed by Sutton *et al.* [35] where participants were completing scores of valence-arousal-dominance using SAMs for 1363 different images. Then, the ratings were compared using the existing International Affective Picture System and the ANEW. The idea was to increase the amount of control in studies examining the perception, processing and identification of facial expressions.

For utilizing the VAE in this investigation, some research was done to see how well it performs on emotion recognition tasks. Li *et al.* [22] tried to utilize a variational autoencoder (VAE) to determine the latent space within a multichannel EEG. They have performed this approach on the DEAP and SEED datasets, two public datasets used for this type of purposes, and also compared VAE with other autoencoder-based (AE) approach as well as Independent Component Analysis (ICA). The author states that this is the first work that introduces VAEs for decoding multichannel EEG. After the signals' decoding contextual learning has been performed, more specifically the Long-Short term memory (LSTM) network to perform sequence modelling for emotion prediction. Again, this approach was compared with some additional baseline methods such as SVM, RFs, K-NNs, logistic regression, Naïve Bayes and a feed-forward deep neural network. The metrics that have been used to compare the performance between models, were the Pearson correlation coefficient that was used to determine the difference between the original input from the multi-channel EEG and the reconstructed one. Overall, this approach is a robust technique for the reconstruction of a time series (EEG) as stated in the paper. For the performance of emotion predictions, they took data only from one participant as their test data the rest were cross validation was performed on the rest (training data). This was done to compare their framework with the baseline approaches. The metric used to model performance was the F1-score. For the final evaluation for EEG reconstruction, the VAE stabilised at a 0.9 correlation coefficient on the DEAP dataset. On this dataset, the AE and RBM obtained approximately 1.0 correlation on this dataset while increasing the number of latent factors. Although, on the SEED dataset the VAE seems to perform much better than the rest while the latent factors were fewer. It implies that the VAE can extract the most relevant factors, regarding emotions from a multichannel EEG. Although, when evaluating performance for emotion prediction with LSTM based on all the reconstruction approaches mentioned above the VAE+LSTM achieved the highest F1-score of 0.8429 and ICA+LSTM achieved 0.6994 for modelling 31 latent factors on the SEED dataset. Also, on the DEAP dataset again the VAE+LSTM achieved the highest F1-scores of 0.7167 (Valence) and 0.7243 (Arousal) for modelling 16 latent factors.

For the fact of generating labels for an 'unlabelled' domain, Zhang *et al.* [43], proposed a semi-supervised learning approach due to how time consuming and expensive EEG annotation is. Although, we do it because we can't really trust the SAM and the pre-determined label on the ANEW. After doing some pre-processing on the raw signals, they have extracted the differential entropy features yielding 310 features in total. They then, used deep recurrent autoencoders to model the latent space from the DE features of the EEG. They have also utilised the attention mechanism. Although, they have used a classifier on the bottle-neck layer(latent representation) of the AE, where here the supervised learning method takes place. The loss function used for the reconstruction of the signals on both those that were labelled and unlabelled was the sum of the MSE losses for both cases. The loss function used for the classifier was cross-entropy. Implementation of this was performed on the SEED dataset. Following some results that were obtained from their approach, in comparison with work from other researchers, utilizing the Attention Recurrent AE have achieved the highest accuracy

with respect to the rest of 91.17%. Emotion recognition with EEG becomes even broader through the years.

Many questions have already been answered but also there are a lot of factors that are still under investigation. Li *et al.* [20] introduced the idea of noisy labels, where they may arise through many factors that humans can cause. This may include natural biases, subjectiveness or even inconsistencies in their judgement of emotions. They have proposed a classification method using a capsule network (JO-CapsNet) and pseudo labels are updated by predicting the output class label based on the network.

All the above are sufficient enough solutions to be implemented for medical diagnoses of mental illnesses. They keep expanding and also certain achievements are being verified. Li *et al.* [22] stated that their work is very promising solution to also diagnose mental illnesses such as depression, Alzheimer's disease, mild cognitive impairement etc. Alchalabi *et al.* [1] investigated the integration of an EEG-controlled serious game that trains and strengthens patients' attention. For this problem, they have achieved a classification accuracy of 96% for correctly detecting the attention state while gameplaying. This is a proof that can be used in the detection of patients suffering from ADHD.

2.5 Hypotheses

(H1) As noted by Li *et al.* [22], they report that performing EEG decoding with a VAE yields better F1-scores instead of implementing ICA. Although, they obtained results for implementing classification models. Instead, as we mentioned in chapter 1 implementing regression models will give more precise predictions and can be more generalised in terms of including all possible emotions. In practice, performing ICA in addition with VAE for the preprocessing pipeline achieves different results in regression than ICA itself.

(H2) Visualising EEG emotion recognition analysis with the inclusion of 'pseudo' labelling as discussed by Li *et al.* [20], makes the problem more realistic due to the inconsistencies that participant can bring while collecting EEG data with emotion scores. Instead, from the findings of Li *et al.* [22] that the VAE achieves brilliant results, it can be furtherly explored in a semi-supervised learning framework by implementing 'pseudo' labelling for the incorrectly labelled scores. This approach can only be conducted for the dataset that we acquired data ourselves.

Chapter 3

Methodology

This chapter will breakdown the pipeline that is implemented for the experiment. Firstly, data acquisition will be explained in terms of what equipment is used and the experimental paradigm used to collect data from participants. Then some pre-processing techniques, to remove unnecessary artifacts, the feature extraction pipeline to determine the latent features that we look for within EEG signals, and lastly the regression models used to predict the VAD scores. All the code written for this project can be found in the following github link (`https://github.com/zazass8/MSc-Dissertation`).

3.1 Data acquisition

For data collection 12 participants have been chosen, from the University of Bath. Out of 12 participants, 11 were males and 1 was a female. They are ageing between $22 - 25$ years old. An amount of 36 images has been selected from the UCDSEE database [41]. All 36 images elicit 9 different emotions, including anger, disgust, embarrassment, fear, happiness, pride, sadness, shame, and surprise where each emotion was elicited by 4 pictures each. Furthermore, a SAM has been given to each participant. It was used to rate their own scores of valence-arousal-dominance after their experience with each image.

3.1.1 Equipment

The EEG device that was used for these experiments is the 16 channel EMOTIV EPOC dry-contact headset, also shown in figure 3.1. The headset it designed in a manner that can cover the whole scalp of the participant, particularly the 14 channels of desire were the F3, FC5, AF3, F7, T7, P7, O1, O2, P8, T8, F8, AF4, FC6 and F4. A figure as in 3.2 shows the locations of these channels. Signals were acquired with a 128Hz sampling frequency, and a notch filter is already planted to the headset to narrow down the signal even more. We need to pay particular attention to those adjusted on the frontal lobe. Mainly those are responsible for detecting emotions, in terms of channels these are the F3, F4 and F8. Figure 3.3 illustrates which part of the brain the frontal lobe is located. Furthermore, a table of 4 columns on an A4 paper was used to design the SAM where the first column was the stimulus index, the second was valence, third was arousal and fourth was dominance. Pens were provided as well so that participants could right their scores of the SAM.

Figure 3.1: 16 channel EMOTIV EPOC dry-contact headset used for data acquisition of EEG signals. 14 of those electrodes are used to gather data and the last 2 as reference electrodes.

3.1.2 Procedure

The experimental paradigm used for this investigation was inspired from Sutton [35] *et al.*. The recordings will take place in a room where only a limited amount of light can surpass as well as that is completely empty of any electrical appliances near the participant's chair. We also made sure that any electrical appliances and plugs are switched off to avoid any interference of external signals. Firstly, we tried and introduce the experimentation that took place and guide the participant about the meaning of the scores of VAD, how the scales of those scores work. The scales were starting from 1 meaning 'highly negative' up to 9 being 'highly positive' and 5 being 'neutral' for all 3 measurements. Further guidance was also given with figures 1.1 and 1.2. They were instructed that they were going to visualise a picture of a person making a facial expression for 4 seconds without making any muscle/eye movements, then take 10 seconds to complete the SAM for that picture and then another 5 seconds to rest and try to stabilise themselves to the screen and prepare for the next one. The reason rest was given was because the device is very sensitive to minor movements, leading to a decrease in conductivity with the participant's scalp. We needed to make sure that there was 100% conductivity while visualising pictures. They were instructed that they have to rate each image based on their immediate personal reaction as well as not comparing any of the images to each other and to rate them individually [35]. This was repeated for 108 images, mainly using the UCDSEE database 3 times but on the second and third trial, images were shuffled with respect to the previous order. Participants will be informed before the experiment that there is no right or wrong answer on whatever they believe about their scores. Before starting the procedure, they were given a practice sample from the WSEFEP database [28] so that they can get familiar with the task and with the scales. That also helped to ensure if participants needed more time in the SAM and REST tasks. Including the adjustments of the time intervals, the whole experiment for each participant took approximately between $37 - 53$ minutes. With that we also made sure that participants weren't getting too tired through the process and were completely focused. Some pictures of this procedure in appendix A.

3.1.3 The DEAP dataset

As we are using our data to check preprocessing and machine learning pipelines, we also use the DEAP dataset just for a sanity check that the models work well as well as comparing results with our data. The DEAP database was used for emotion analysis using physiological signals. The researchers collected data from 32 participant, where each participant experienced

40 one-minute long videos. Similarly participant were rating each video on a SAM in terms of valence, arousal, dominance as well as liking and familiarity with the stimuli. We will only compare scores of valence, arousal and dominance instead. While each stimulus was elicited, researchers collected EEG data for using 32 channels at a 512Hz sampling frequency. The metrics that were used were accuracy and F1-score as they performed classification for machine learning [17]. Although, we need to compare results based on RMSE and MAE hence for a good comparison, we'll re-do the machine learning to get those metrics and compare more efficiently.

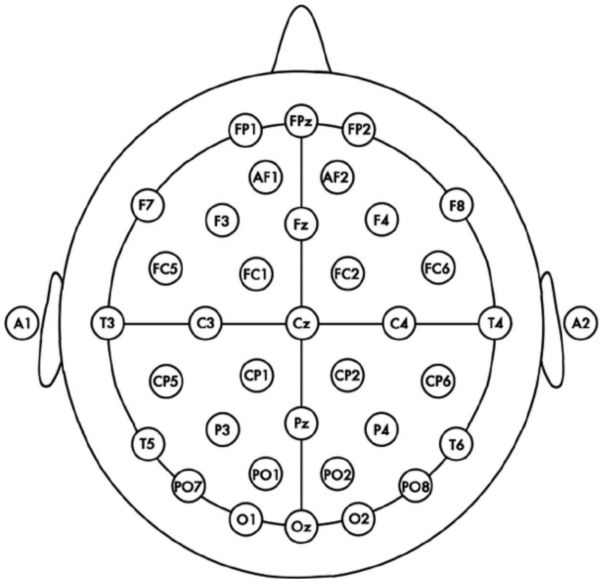

Figure 3.2: 37 channel montage from [5] showing analytically the location of each electrode of the device. All 16 channels that we used are already included.

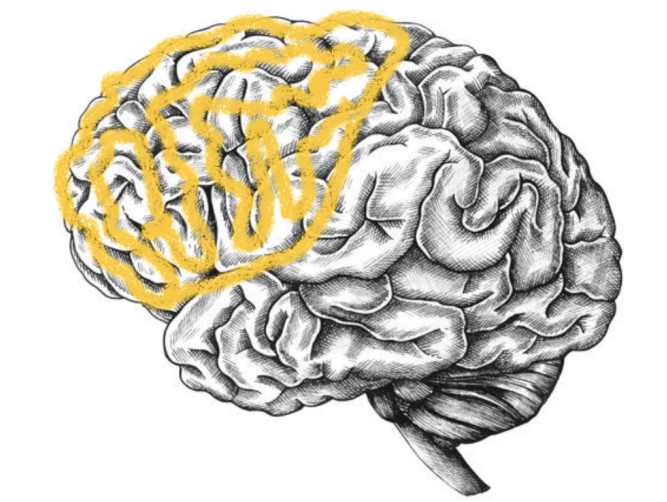

Figure 3.3: Human brain structure. Clearly shown with orange colour the region of the frontal lobe which is of interest.

3.2 Preprocessing

It is known that brain signals have a very abstract structure, containing many artifacts and noise. These may appear due to external factors while gathering data, such as heartbeat (ECG), respiration, muscle movements (EMG) or even blinking of the eyes (EOG) [30]. Although, there are types of artifacts such as enviromental (power line noise, door slamming, electromagnetic field noise) and instrumentation (malfunctions of the EEG device) that can't be really controlled and are really difficult to clean out from the EEG signal. Before conducting any experiment with participants, we made sure we followed the precautions as mentioned in section 3.1.2. The room where experiments took place was dim enough, electronic and mobile devices were switched off.

From a machine learning framework this can be thought of a dimensionality reduction problem. In the following sections, we'll describe the algorithms that are used. But before executing any of these algorithms, the data must be filtered first using a band-pass filter. The reason is because only certain frequency bands will be useful regarding the detection of emotions and these are the θ (4-7 Hz), α (8-13 Hz), β (13-30 Hz) and γ (>30 Hz) bands [9]. More certainly, high frequency waves like β and γ are used to correlate arousal, low frequency waves like α and θ are used for valence and all these contribute in correlating dominance as well [36]. These bands are between the 4-50 Hz frequency domain and the bandpass filter is adjusted to filter out the rest. Furthermore, artifacts like power line noise which occurs in the 50-60Hz domain, are already filtered out by the EPOC device as it has a notch filter planted to it relatively for this purpose. The EPOC samples data with a 128Hz sampling frequency and as already mentioned, we use a band-pass filter to retain only the frequencies between 4-50Hz. That also ensures that the EMG signals are filtered out as those belong to the domain of frequencies above 50Hz [32].

Here's an illustration in figure 3.4 of how a raw brain signal looks like. These are 48 channel data from a single participant for the first 20 epochs. We are informed from the dataset description in [27] that channels from EXG1 up to EXG4 are EOG channels and channels from EXG5 up to EXG8 are EMG. To filter it out and remove the EOG signals, independent component analysis (ICA) will be conducted as explained in section 3.2.1 and additionally a variational autoencoder (VAE) as in section 3.2.2 to compress the already preprocessed signal while determining its latent representation that contains the most information of it. These two approaches are going to be compared.

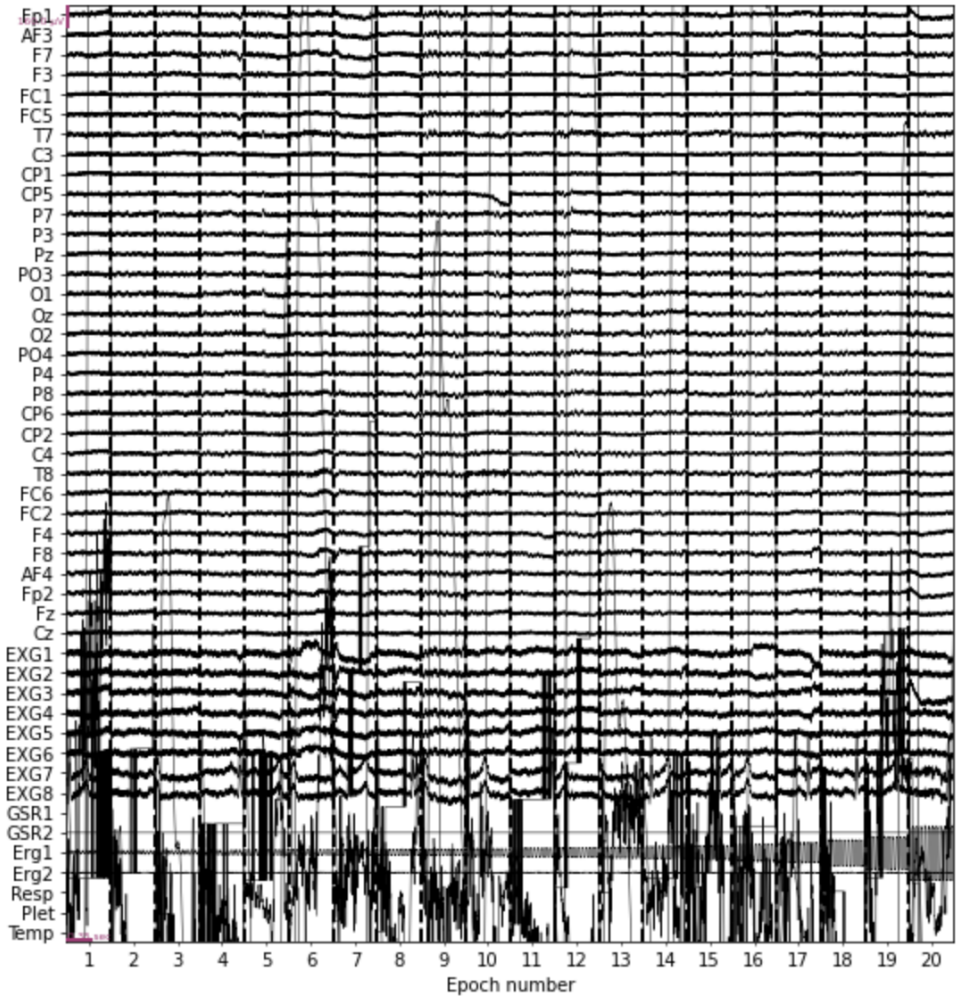

Figure 3.4: Raw signal from DEAP dataset, showing clearly the noisy channels from EXG1 up to Temp.

3.2.1 Independent Component Analysis

Independent Component Analysis (ICA) is considered the standard approach in the domain of signal processing, specifically on the 'denoising' of signals. Imagine a mixture of signals that are generated from several sources. What the algorithm does, it unmixes these signals and detects these independent sources in other words from where do signal originally come from. With reference to the cocktail-party problem [26], ICA is considered a great solution. Imagine a cocktail party takes place in a room and several microphones are situated in different location within the room. Assume there are external sources of sound within the room, such as loud music, people chanting, noise from the roads etc. These sounds are considered 'source' signals and all microphones acquire all these sounds. ICA is capable to extract the original sound sources from these sound mixtures and making them independent from each other.

With EEG, it works really good for data cleaning more specifically removal of artifacts within the EEG data. An example of how these can be detected, is if there is a large amplitude at certain frequencies within the signals. And what the algorithm does, it decomposes them by maximally finding these statistically independent sources of variance and uses that to exclude the ones with the large amplitudes [29].

In a more mathematical formulation suppose that $\mathbf{X}$ is the raw signal that is already noisy. Then with a matrix multiplication this problem is formulated as

$$\mathbf{S} = \mathbf{W}^{-1}\mathbf{X} \tag{3.1}$$

where $\mathbf{U}$ is the matrix with the ICA source activities, $\mathbf{W}^{-1}$ is the unmixing matrix that is the one we are trying to learn and $\mathbf{X}$ is the raw signal. Hence, ICA attempts to make each row of U statistically independent with no constraint on W. And to prove that $\mathbf{X}$ is noisy, it can be formulated as

$$\mathbf{X} = \mathbf{WS} \tag{3.2}$$

where $\mathbf{W}$ is the mixing matrix. Although, the main difference with the PCA algorithm is that PCA decorrelates $\mathbf{U}$ assuming that $\mathbf{W}$ is orthogonal, whereas ICA attempts to make $\mathbf{U}$ statistically independent with no constraints on $\mathbf{W}$. In that sense, using singular value decomposition $\mathbf{W}$ is decomposed as

$$\mathbf{W} = \mathbf{U\Sigma V^T} \tag{3.3}$$

where $\mathbf{U} = \mathbf{W^T W}$ and $\mathbf{V} = \mathbf{W W^T}$ and $\mathbf{\Sigma}$ is the noise covariance matrix. Inverting $\mathbf{W}$ will give

$$\mathbf{W^{-1}} = \mathbf{V\Sigma^{-1}U^T} \tag{3.4}$$

So before conducting the actual ICA method, using these information data are first whitened using the PCA algorithm. The problem in equation **??** is reformulated as

$$\mathbf{S} = \mathbf{VX_w} \tag{3.5}$$

where $\mathbf{X_w}$ is a whitening transformation of the data which means it transforms a vector such that its components become uncorrelated. This leads to a much easier problem and we can learn $\mathbf{X_w}$ with maximum likelihood estimation [39]. Hence, for the decoding of multichannel EEG signals ICA is used to unmix the signal into its independent components which also makes it easier to detect the EOG and ECG artifacts. There are a lot of algorithms that can do ICA and the one we use is FastICA. The reason this algorithm is used is because, it converges much faster than the rest as well as there is no need to adjust many parameters like learning rate etc which also makes it more user friendly [39].

The way we apply ICA on EEG data is with the MNE library of Python, where we generate several plots that aid in judging correctly which independent components needs to be excluded, in our case the EOG channels. An illustration of how independent components look like, once we fit it with the data is given in figure 3.5. With MNE, out of 40 ICs, 19 were only visualised. Although, this plot is for the sake of an additional visualisation, a more correct decision making is done by visualising topographic heatmaps for each channel. A plot of heatmaps in figure 3.6 describes for all ICs at which regions of the participant's scalp the EPOC device gathers information. The red colour indicated that there is a lot of information within that region, and is it gets blue only limited information exists. This time, we cropped out the first 14 ICs out of 40.

And more specifically, we can visualise further plots to make a more efficient judgement for each component. In figures 3.7 and 3.8 we can see several plots regarding independent components individually from a single participant but from the data that we collected ourselves. We decided for this because, the EOG channels can be understood more easily. As you can see, the red regions exist at the front part of the scalp and it only makes sense as it is where

the eyes are located. Hence, those channels are going to gather eye blinking artifacts more easily. The top-right plots illustrated how noisy independent components are. The more red the heatmap appears, it depicts strong influence within components. Hence, the more noise it contains. It does not necessarily mean that wherever red dots appear, it means that those components should be removed.

More ideally, it is much better to leave some noise rather than removing pure signal. Furthermore, we may remove artifacts in terms of their overall contribution to the overall signal. This can be assessed based on the degree of variance within it and it is illustrated on the bottom right plots. The more sparse the distribution of data is, the more influential it is. And finally, the bottom left plots, illustrate the power spectrum of components. On those plots, artifacts are detected in terms of spikes that appear at certain frequencies. In those case, applying notch filtering to those frequencies can easily get rid of artifacts. More details on power spectral density in section 3.3. The plots for the signals before and after ICA are shown in figures 3.9 and 3.10, where it is clearly illustrated that ICA has 'cleaned' the signal efficiently.

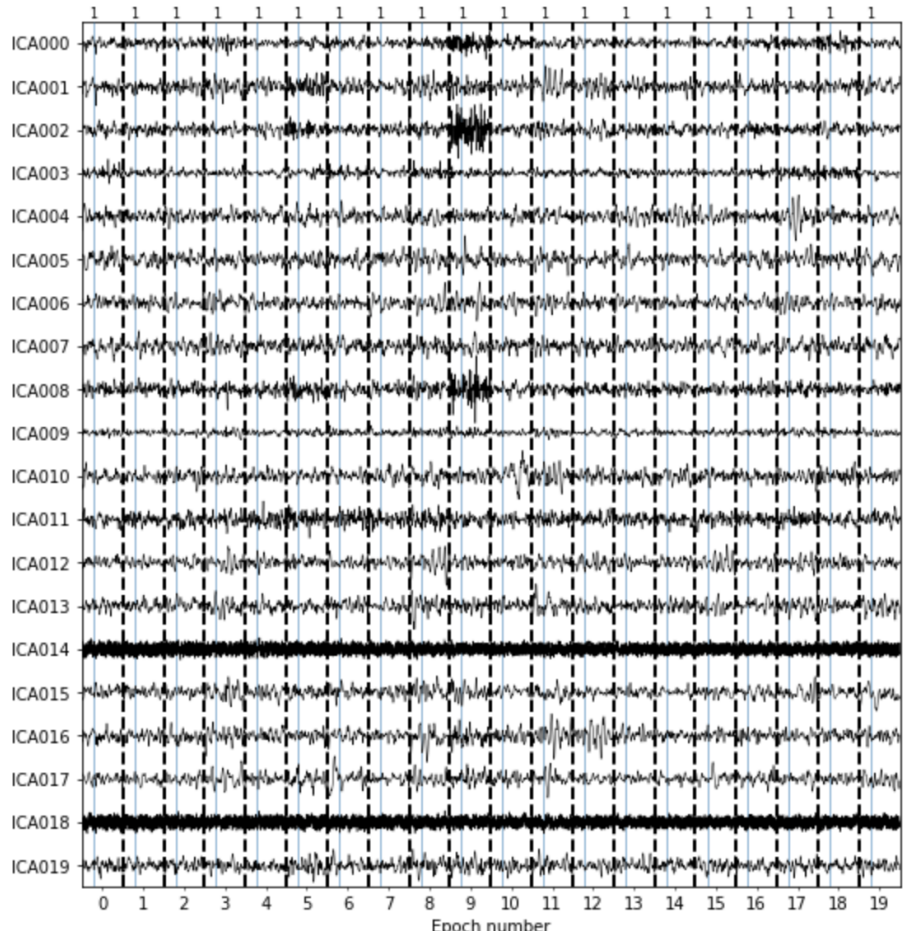

Figure 3.5: Independent components when applying ICA to the raw data. Signals are well separated making it easier to detect the signals with artifacts.

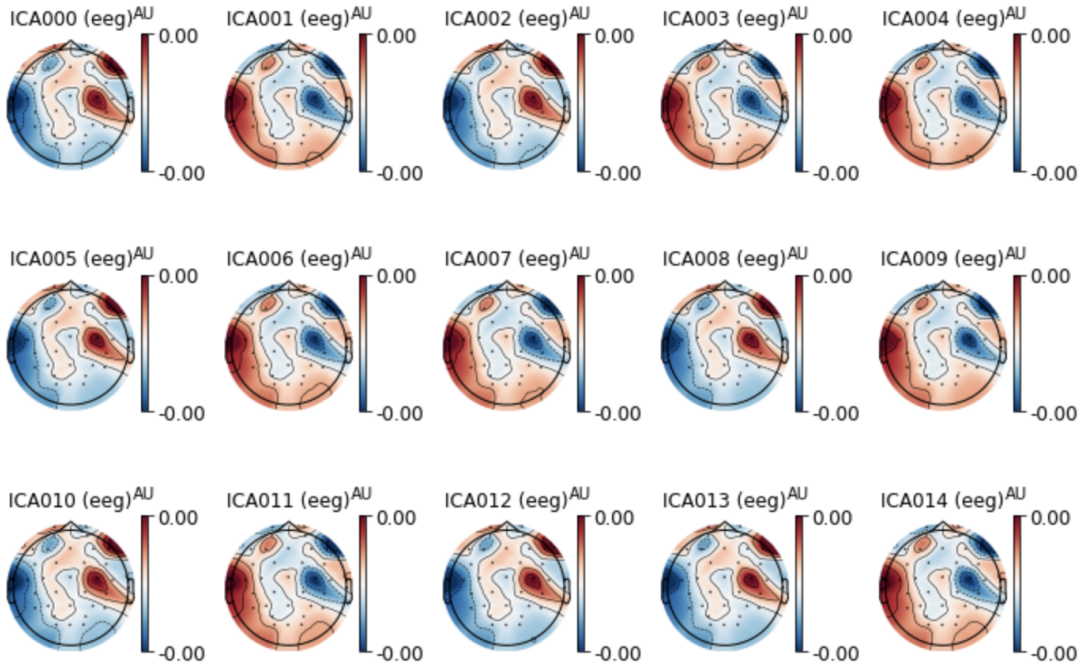

Figure 3.6: Topographic heatmaps of all independent components. Each indicates with the red regions at which part of the scalp the Epoc device gathers data.

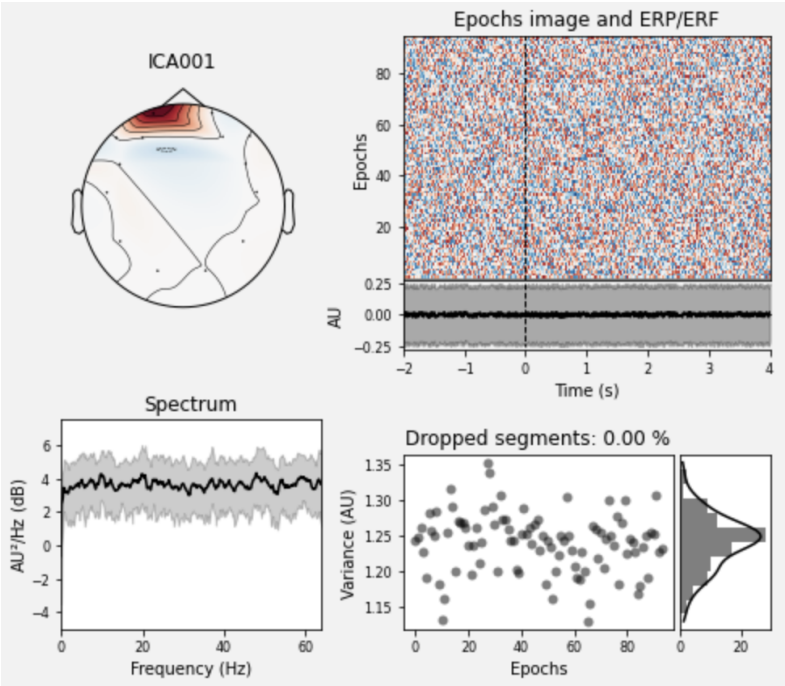

Figure 3.7: EOG component 1, with top left indicating which part of the scalp the Epoc device gathers data and top right showing that more red spots there are, the more noisy independent components are. Bottom left shows the power spectral density, observing noise with spikes and at which frequencies and bottom right showing the degree of variance within it and it is illustrated on the bottom right plots. The more sparse the distribution of data is, the more influential it is.

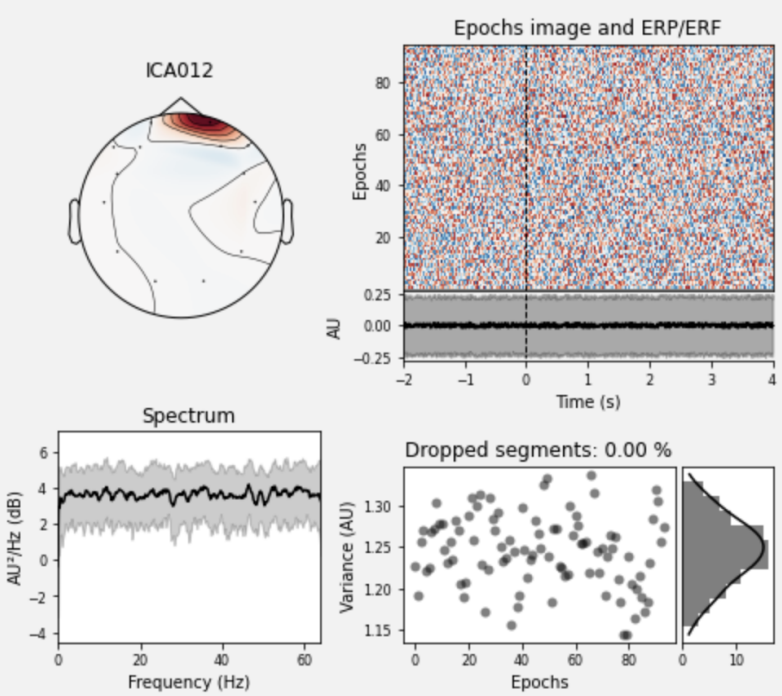

Figure 3.8: EOG component 2, with top left indicating which part of the scalp the Epoc device gathers data and top right showing that more red spots there are, the more noisy independent components are. Bottom left shows the power spectral density, observing noise with spikes and at which frequencies and bottom right showing the degree of variance within it and it is illustrated on the bottom right plots. The more sparse the distribution of data is, the more influential it is.

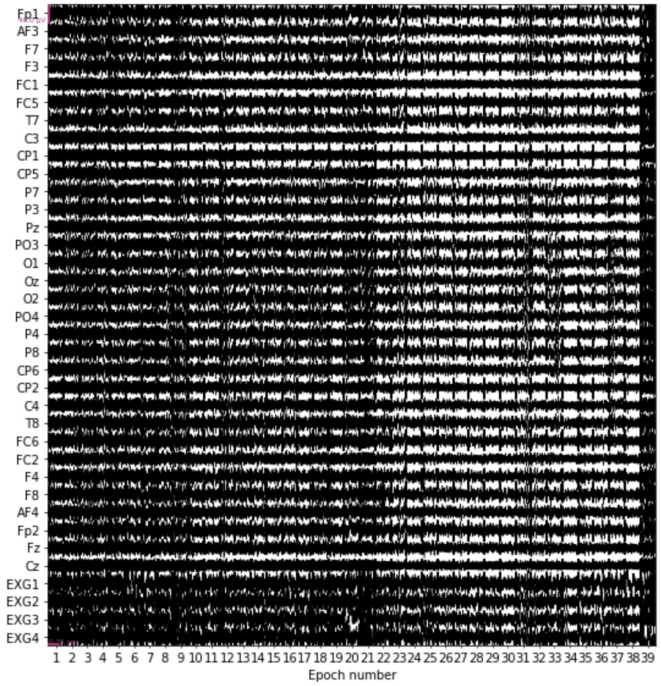

Figure 3.9: Raw signal before ICA. Can be shown easily how noisy it looks before applying the algorithm.

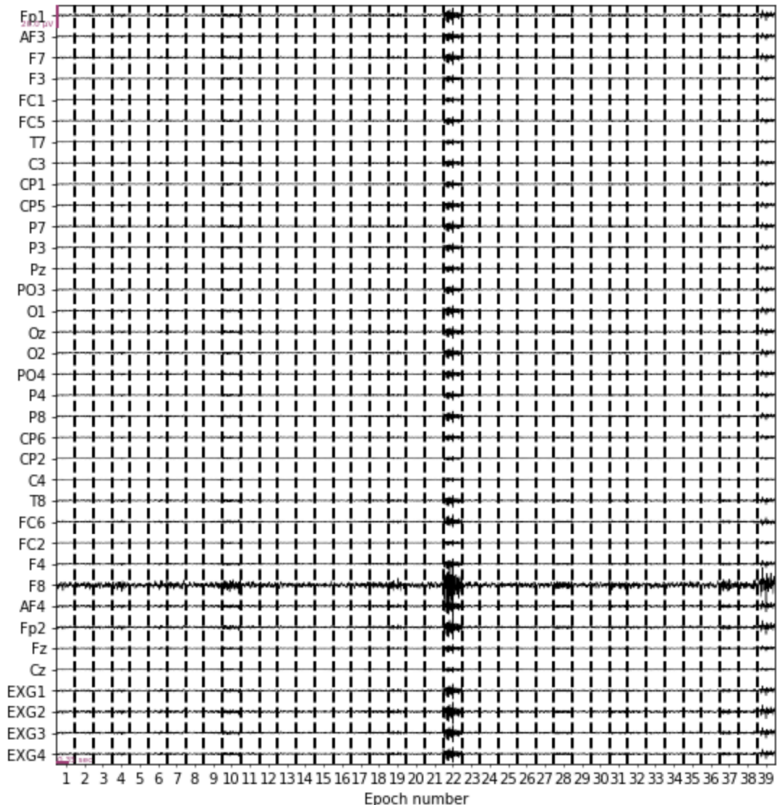

Figure 3.10: Raw signal after ICA. More clear representation of signals with artifacts already excluded.

3.2.2 Variational Autoencoder

From a machine learning framework this can be thought of a dimensionality reduction problem. We have large datasets with many unnecessary features, and we try to extract many of them but only preserving the ones with the most information. Through the literature there are various machine learning techniques to deal with this. One can go with traditional unsupervised learning approaches such as Principal Component Analysis (PCA) or ICA. There is a more robust technique such as the implementation of an autoencoder (AE). Autoencoders are considered an unsupervised learning technique as well and after their training, the part of the encoder only remains where it would be used to compress the original input (from EEG). As shown in figure 3.11, the original input is compressed (encoder), then it tries to reconstruct the input (decoder). Although the reconstructed input is a prediction of the original input, backpropagation is performed to minimise the difference between the original and the reconstructed. After the network is trained, the decoder is removed. Various types of autoencoders such as Long Short term memory autoencoder (LSTM-AE), restricted Boltzmann machine (RBM) or VAE could be used [22]. But for a purpose like ours that we are trying to pursue, and also according to the findings of Li [22] *et al.* the VAE could be the best option. The only difference that a VAE has from a standard AE is that it is formulated as a density estimation problem. The assumption is that all data are generated by a random process that involves a latent variable z. This variable is generated from a prior distribution $p_\theta(z)$ where θ is unknown as well. The input is formulated as $p_\theta(x|z)$. The posterior distribution then is formulated as:

$$p_\theta(z|x) \propto p_\theta(x|z)p_\theta(z) \tag{3.6}$$

Directly computing the posterior is intractable, hence a proposal distribution $q_\phi(z|x)$ is introduced to approximate the posterior. It is assumed that $q_\phi(z|x)$ and $p_\theta(z)$ follow a multivariate Gaussian distribution. Hence, in VAE the encoder encodes the input in latent variables $q_\phi(z|x)$ and the decoder maps these variables to reconstructed input $p_\theta(x|z)$. In other words, the autoencoder tries to learn the latent mean and variance of the gaussian distribution and then draws a sample for a latent vector from it. The optimization problem is formulated as:

$$\max \left\{ E_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p_\theta(z)) \right\} \tag{3.7}$$

where the above term is known as the variational lower bound [22]. The left term is mainly the loss for the reconstruction of the signal and the right term is a regularisation loss. Additionally, a representation of its architecture is also illustrated in figure 3.12.

So due to how different a structure of a brain signal can be, a generative model like the VAE could be more useful to model the hidden state space of the brain [22]. And more specifically, the usefulness or reconstructing the compressed signal back to the original input, helps understanding how noisy the compressed input is. Ideally, the closer the reconstructed input is to the original one, the less noise the compressed signal contains. But of course, because deep learning is implemented for this task it would be good to apply an approach like ICA just for the comparisons between deep learning and general unsupervised learning, hence also proving that AEs perform better. Using `keras` from `tensorflow`, we can visualise the architectures of the neural networks used for training and it is illustrated in figure 3.13 for our dataset, with its analytic summary given in figure 3.14.

For the DEAP dataset it is similar, with only difference the numbers of trainable parameters on each layer's input and output. It is clearly shown from the architecture that 2D convolutional layers have been added in both the encoder and decoder with Leaky ReLU as activations. Convolutional layers are used to drop the dimensionality of the given input and preserving the most important information by applying the convolution operation. Unlike ReLU, Leaky ReLU has a small slope for negative values of the input and it helps in handling sparse gradients during backpropagation. They are highly recommended for the training of deep generative models such as VAEs. Furthermore, that there have been included some Batch Normalization and Dropout layers. These are included to avoid the effect of the vanishing gradient in backpropagation and overfitting. For training, we used RMSPROP as our optimiser [22] with a learning rate of 0.0001. It is observed that the lower the learning rate, the more efficient the training of the VAE. A separate batch size of 2 for the DEAP and 5 for our dataset was used for training, in order to be divisible with the number of datapoints that undergo training.

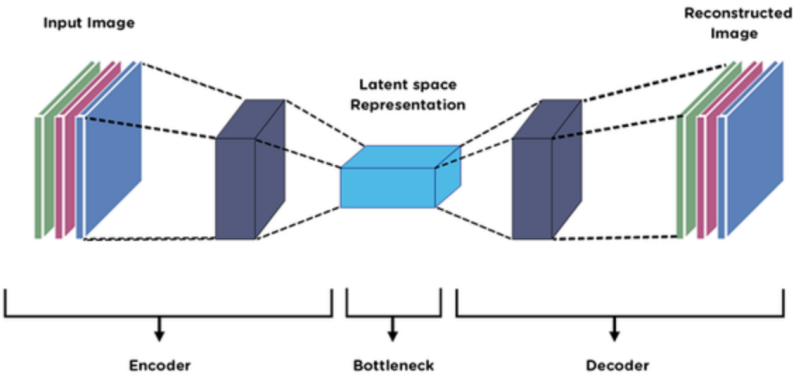

Figure 3.11: Traditional autoencoder architecture from [3]. The input signal is fed to the encoder and then it is compressed in the bottleneck layer. After that, it is decompressed and reconstructed back to its original formation once fed from bottleneck to the decoder.

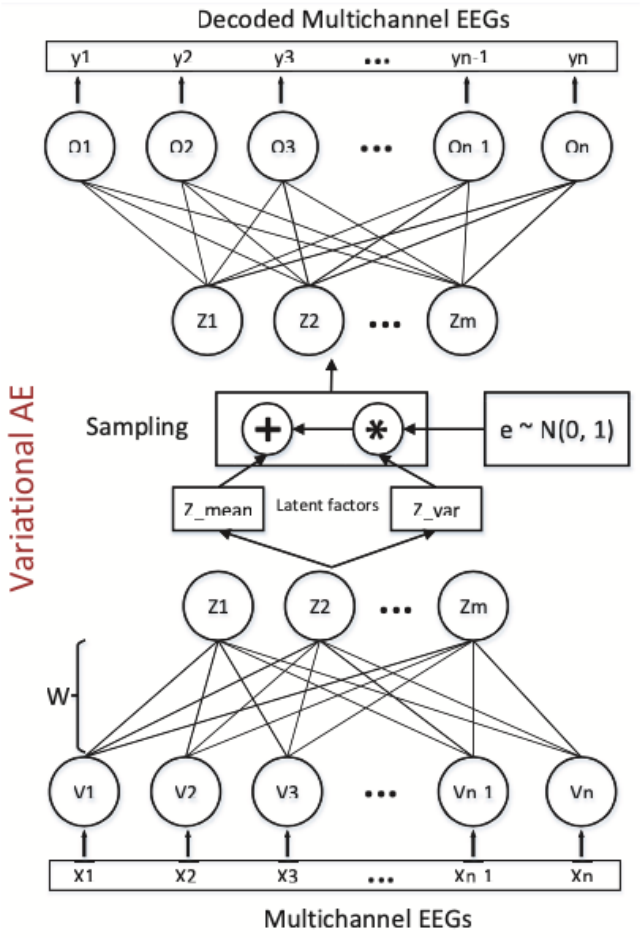

Figure 3.12: VAE architecture from [21]. Learning the latent mean (z_{mean}) and variance (z_{var}) and applying the reparametrisation trick of $z = \mu + \sigma \odot \epsilon$ to sample a latent vector z directly to feed the decoder.

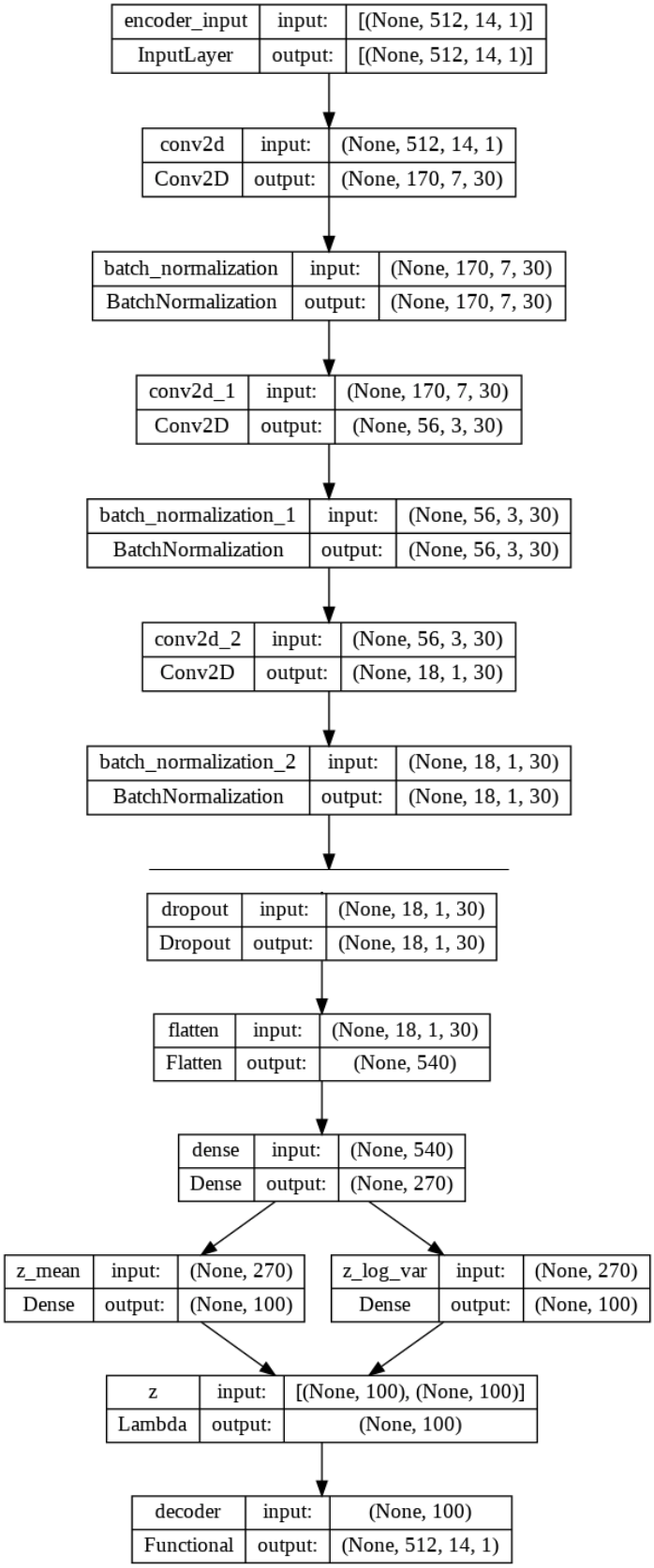

Figure 3.13: VAE architecture for our dataset, showing analytically the dimensions of signals before and after applying layers.

Layer (type)	Output Shape	Param #	Connected to
encoder_input (InputLayer)	[(None, 512, 14, 1)]	0	[]
conv2d (Conv2D)	(None, 170, 7, 30)	210	['encoder_input[0][0]']
batch_normalization (BatchNormalization)	(None, 170, 7, 30)	120	['conv2d[0][0]']
conv2d_1 (Conv2D)	(None, 56, 3, 30)	5430	['batch_normalization[0][0]']
batch_normalization_1 (BatchNormalization)	(None, 56, 3, 30)	120	['conv2d_1[0][0]']
conv2d_2 (Conv2D)	(None, 18, 1, 30)	5430	['batch_normalization_1[0][0]']
batch_normalization_2 (BatchNormalization)	(None, 18, 1, 30)	120	['conv2d_2[0][0]']
dropout (Dropout)	(None, 18, 1, 30)	0	['batch_normalization_2[0][0]']
flatten (Flatten)	(None, 540)	0	['dropout[0][0]']
dense (Dense)	(None, 270)	146070	['flatten[0][0]']
z_mean (Dense)	(None, 100)	27100	['dense[0][0]']
z_log_var (Dense)	(None, 100)	27100	['dense[0][0]']
z (Lambda)	(None, 100)	0	['z_mean[0][0]', 'z_log_var[0][0]']
decoder (Functional)	(None, 512, 14, 1)	184895	['z[0][0]']

Figure 3.14: VAE architecture summary of layers and trainable parameters for our dataset. 396,595 parameters are used for training and 302 are non-trainable due to the inclusion of dropout layers

3.3 Feature extraction

When an EEG signal is decoded and artifacts are completely removed from it, it can be passed to the machine learning pipeline. As mentioned in section 3.2, we desire to gather information about emotions from the θ, α, β and γ frequencies which is already done with the band-pass filter. But, the signal needs further decomposition instead so that the machine learning model can learn from it more efficiently. There are a lot of methods of performing this, such as FFT, STFT or Wavelet Transform. Or further decomposing the signal, and determining features that describe better valence or arousal, for example power spectral density (PSD), fractal dimension (FD) and discrete wavelet transform (DWT) work best to determine valence where FD and DWT work best to determine arousal [37].

Furthermore, several statistical features are considered when analysing EEG signals. These may include features like, minimum, maximum, mean, standard deviation, skewness, and kurtosis. They are useful in a sense that they can describe the characteristics of a signal. For example, minimum and maximum can describe how large the amplitudes of the signals are. Skewness and kurtosis, indicate the largest dispersion and are useful in the evaluation of different mental states [14].

From all these, we chose PSD for feature extraction. A signal is broken down into a weighted sum of sinusoidal waves. Each signal is composed of three features and these are frequency, amplitude and phase. The power spectral density disregards the phase feature and it is certainly described as a weighted sum of all the amplitudes of those sinusoidal waves for each possible frequency of that signal. Hence, it describes better the relative importance of each frequency

component to the overall signal [34]. This function can be thought of as a probability density function but instead the density depicts power spectrum for each possible frequency. The reason this method is chosen, is because it perfectly can represent the EEG data in the frequency domain, by windowing them into bands as well as computing the weighted sum of all amplitudes for each frequency makes each component of the signal unique. Hence, the machine learning model can learn more efficiently from it. To compute it we used the Welch algorithm and what the algorithm does, it partitions the data in K batches. Then, for each segment it computes a windowed discrete Fourier transform and it is defined as

$$X_k(\nu) = \sum_m x_m w_m e^{-2\pi i \nu m} \tag{3.8}$$

where ν is the frequency and w_m is the window function. From here, we form the modified periodogram value which is defined as

$$P_k(\nu) = \frac{1}{\sum_m^M w_m^2} |X_k(\nu)|^2 \tag{3.9}$$

Then, we average the periodogram values to obtain the PSD as

$$S_x(\nu) = \frac{1}{K} \sum_k^K P_k(\nu) \tag{3.10}$$

3.4 Pseudo-Labelling

As mentioned in chapter 1, we cannot really trust individual recordings from participants and we additionally use the ANEW to sanity-check and compare the results based on that. Then we are combining supervised with unsupervised learning techniques to generate labels via pseudo labelling. Pseudo-labelling is a technique that is used to generate synthetic labels for the unlabelled EEG signals. This is done by training a model on the labelled signals that we have, then using the model that was trained we will generate synthetic labels on the unlabelled data. Finally, we will augment the signals with generated label to the original dataset of the labelled signals and then re-train the model [19]. For the signal processing approach that utilises a VAE, the loss function that is used is similar as in equation 3.7 but with an extra term a shown below

$$\max \left\{ E_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p_\theta(z)) - ||y - \hat{y}||^2 \right\} \tag{3.11}$$

where y is the true label for the data that are already labelled and $\hat{y}$ is its prediction [43]. In EEG emotion recognition, while a participant records the labels of stimuli on the SAM he/she certainly can make mistakes. This is due to humans having natural bias and inconsistencies in their judgement leading to the final labels they assign to be noisy. For example, in the case of visualising facial stimuli, there some ambiguity in certain emotions. This can easily lead to some confusion and the participant can't be really sure what he/she will be writing as ratings. Through several studies, it has been acknowledged that emotions are subjective and overall humans understand and perceive emotions differently [7].

In this investigation, pseudo labels are not generated simultaneously while conducting feature extraction either with ICA or VAE. Hence, the third term in equation 3.11 ($||y - \hat{y}||^2$) is dropped. Additionally further feature extraction using the Welch algorithm is done as described in section 3.3, and only then the data are trained to generate the pseudo labels to the unlabelled

data. The classifier that is used to predict the missing labels is SVM with a rbf kernel and a regularisation of 1.0, as the labels are assumed to be 'noisy' and the data are non-linear as shown in figure 3.15. Finally, in the discussion part in section 4.2 the generated labels are assessed based with the ANEW after doing signal processing both with ICA and VAE.

3.5 Regression

Now, after the first task is completed the next would be the construction of the regression model. Again, there are various approaches for doing regression such as Ensemble learning for Random Forests (RFs), Support Vector Regressors (SVRs), K-Nearest Neighbours (K-NNs), Naïve Bayes or even linear regression. From these techniques, the ones that perform best are mainly SVR's and RF's, while techniques like Naïve Bayes won't be that useful for these types of tasks. Hence, we may proceed in implementing SVRs, RFs, and k-NNs. The metrics that will be used to compare results are mean absolute error (MAE) and root mean square error (RMSE). It is also recommended to perform the 5-fold Cross Validation technique as to have a better indication of accuracy in the training data, leading in a better and overall performance of the model. In python this is achieved with `multioutput.MultiOutputRegressor` from `sklearn` where it simultaneously learns a regression model for all 3 outputs of valence arousal and dominance.

3.5.1 Support Vector Machines

Due to the non-linearity that exist within the data it would be more optimal to use SVRs for the 'kernel trick' [6]. We can also prove it using an example from a single participant's data from the ones we collected. We choose 'anger' and disgust as the two separate classes and we plot all the datapoints that exist within the participant's data in order to check if the data are linearly/non-linearly separable. The plot in figure 3.15 clearly indicates that the data are non-linearly separable and that's the case for all possible classes. Hence, the 'kernel trick' is utilised in order to transform the data into a higher dimensional space and drawing a decision boundary more easily and more effectively. Hence, it does not depend on the dimensionality of the input space [2]. Although, SVR works a bit differently than SVM. It trains a symmetrical loss function which achieves to penalise any high and low estimates. The acceptable range of the ones that are accepted, is defined as the ϵ-tube approach where a symmetric tube of minimal radius is formed around the estimated function. The tube only takes care of the one which achieve an absolute value error below a threshold. The rest are ignored. This approach has also excellent generalisation accuracy with high prediction accuracy [2]. And finally, we decide to add a regularization term as well, as the EEG data that we pass in the model already contain noise, so with it it can adjust better with the effect of 'overfitting'. To visualise how this works, it can be shown in figure 3.16. Three separate models can be proposed with SVR. One would be for the valence dimension, the second for the arousal and the third for dominance [13].

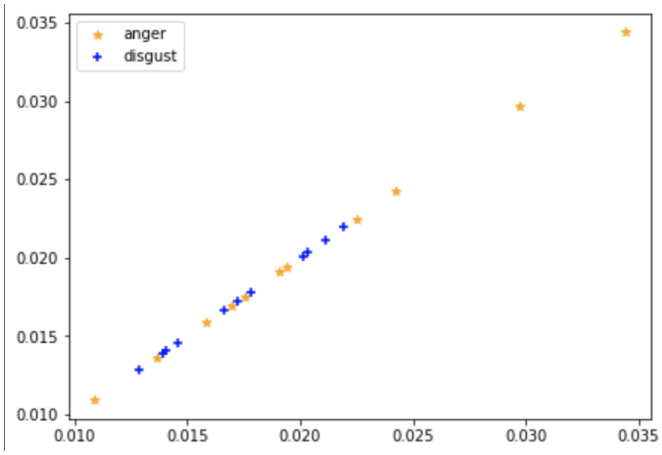

Figure 3.15: Data between 'anger' and 'disgust'. Clearly shown that they are non-linearly separable implying that a non-linear kernel has to be applied.

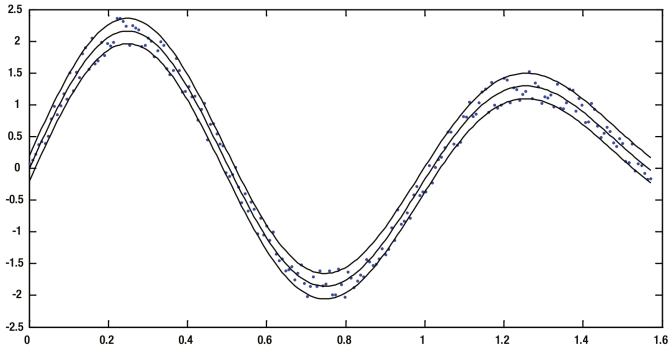

Figure 3.16: Regression with SVR from [2]. The ϵ-tube controls the degree of how large the margins of separations are.

3.5.2 K-Nearest Neighbours

Suppose that we have data for two separate classes and that it is assumed that they are already separated. When a new observation has to be classified between the two classes, the way this is done is by choosing a number (k) of nearest neighbours that this observation falls within. If the observation is nearer to other observation of one class, the it should be classified to that particular class and vice versa. The metric that is used to measure how close data are between the new observations is the Euclidean distance. In regression, for predictions of unobserved data to be drawn effectively the average of the selected nearest neighbours is computed [25]. As in all machine learning algorithms, if the number of k is small enough the model will memorise the training set and will perform badly on the test set, hence it will overfit. On the other hand, if the number of k is large enough, it gets smoother and it underfits. To solve this, we try to adjust k based on this trade-off that occurs between separate models and it is achieved with hyperparameter tuning as explained in section 4.2.

3.5.3 Random Forests

Decision trees is a simple algorithm where they are useful mainly if the data do not follow a regular pattern. Classifications are made, by splitting the data into partitions and the class prediction is made based on the majority of data that exist within that region. Each

split is chosen based on a loss function and in classification trees it's either Gini impurity or information gain. Although, for regression instead of assigning data into classes, the average of that region is used as the prediction [15]. Overall the decision tree tends to overfit and it is not recommended to be utilised. It will definitely overfit in our case with EEG data as they would still contain noise. Meaning, that if the model tends to overfit predictions will have high variance as noise of EEG will be trained as well. Instead, a more advanced version of it has been developed and it's called random forests (RFs). It is a combimation of decision trees with an ensemble method called 'Bagging'. Bagging is a resampling method that is used to shuffle a training set and change its order of data [15]. In this case, n regression trees are trained for n separate training sets because we resample the original training set n times. Then from those n predictions we get, we average them and we get the final prediction of the random forest model. This improves results that decision trees achieve because the variance of predictions decreases hence it does not overfit. For RFs, using hyperparameter tuning we will try and find optimal parameters such as number of decision trees, maximum depth of each decision tree and which loss function is best for choosing a split. More on this in section 4.2.

Chapter 4

Results

In this chapter, all results and evaluations will be inferred in terms of how well the models for feature extraction, regression and removal of noise. The metrics that are utilised for each model are stated in each section. Additionally some inferences and comparisons are done. Firstly, we will start by performing data analysis on the data that we collected from the 10 participants. It has to be stated, that there was a lot of delay in the data acquisition task due to several technical issues. Those were issues with the software used to design the experimental paradigm, as well as of the usage of faulty EMOTIV EPOC headsets and sensors. If this was dealt on time, there was a high possibility of collecting samples from more than 10 participants. Another major issue that was encountered is that a mistake was done while designing the paradigm. The experiment was supposed to record EEG signals with experiencing 108 facial stimuli, but accidentally EEG signals from only 95 were recorded. Hence, the whole investigation was conducted with utilising EEG from 95 images.

4.1 Data Analysis

Overall, participants found the task of data acquisition hard and mistakes were easily done in terms of rating the images. But there were 2 exceptions which lead into excluding that data. For the first occasion, the participant felt sleepy which he/she accidentally missed 2-3 images consecutively, hence losing the whole order of how ratings were given. On the second occasion, the participant found really hard to follow the precautions given before the experiment (eg. no moving, no speaking etc.) and was leaving gaps throughout each image, also forcing himself/herself to go back and complete ratings from past images. For that case the experiment stopped in the middle and never repeated again. Hence, from 12 samples only 10 have been utilised.

Actual results given by participants were compared using the ANEW [4]. The ANEW rates 1060 different words with a mean and standard deviation for valence/arousal/dominance. Hence, if participants' results were falling within the range given by the standard deviation they were retained. Otherwise, they were considered as 'noisy' labels.

Overall, out of the 95 images participants found really hard to classify correctly the valence and dominance ratings. For arousal, they did a really good job only with a few exceptions. Results are also shown in table 4.1.

Further analysis was done to observe at which emotions participants found difficulty to classify

correctly. From most participants, it was observed that embarrassment, shame, and surprise were the hardest to classify in terms of valence, sadness and shame for arousal and disgust, embarrassment and shame for dominance. For simplicity and more efficient visualisation, we only kept the two highest counts of emotions from each participant. Results are shown in tables 4.2, 4.3 and 4.4.

Table 4.1: Amount of mis-classifications by each participant for valence, arousal, dominance scores. Clearly shown that fewer mistakes were done when labelling arousal.

	Valence	Arousal	Dominance
Participant 1	44	17	42
Participant 2	39	6	32
Participant 3	35	5	48
Participant 4	37	14	16
Participant 5	50	2	35
Participant 6	38	19	59
Participant 7	47	21	45
Participant 8	39	19	36
Participant 9	59	17	28
Participant 10	41	17	45

Table 4.2: Amount of mis-classifications by each participant for valence scores. Most mistakes were done for the emotions of 'embarrassment' and 'shame'.

	disgust	emb/sment	sadness	shame	surprise
Participant 1	-	10	-	8	-
Participant 2	9	10	-	-	-
Participant 3	-	9	-	-	10
Participant 4	-	10	6	-	-
Participant 5	-	-	11	9	-
Participant 6	-	10	-	-	8
Participant 7	-	8	-	8	8
Participant 8	-	10	-	-	10
Participant 9	11	10	-	-	-
Participant 10	-	10	10	-	-

Table 4.3: Amount of mis-classifications by each participant for arousal scores. Most mistakes were done for the emotions of 'embarrassment' and 'disgust'.

	anger	disgust	emb/sment	fear	sadness	shame
Participant 1	-	-	-	-	9	-
Participant 2	2	-	-	-	2	-
Participant 3	-	-	-	-	3	-
Participant 4	-	-	-	-	-	8
Participant 5	1	-	-	1	-	-
Participant 6	-	4	4	-	8	-
Participant 7	-	9	-	-	7	-
Participant 8	-	-	-	-	6	4
Participant 9	-	-	-	-	-	9
Participant 10	-	-	5	-	-	5

Table 4.4: Amount of mis-classifications by each participant for dominance scores. Most mistakes were done for the emotions of 'shame'.

	anger	disgust	emb/sment	sadness	shame	surprise
Participant 1	-	-	-	-	9	-
Participant 2	7	10	7	-	-	-
Participant 3	11	-	-	-	-	8
Participant 4	-	-	-	-	8	-
Participant 5	8	8	8	-	-	-
Participant 6	-	10	-	-	10	-
Participant 7	11	-	-	-	-	9
Participant 8	-	-	9	7	7	-
Participant 9	-	7	6	-	-	-
Participant 10	-	-	-	10	11	-

4.2 Signal processing pipelines and pseudo labelling

As mentioned in section 3.2.1, we used the FastICA algorithm for the preprocessing of the EEG signals. And as mentioned in section 3.5 the metrics we have used for machine learning were MAE and RMSE where those are going to be compared based on those. The results for all three ML algorithms of section 3.5 on the DEAP dataset are given in figures 4.1, 4.2 and 4.3 and for our data in 4.4, 4.5 and 4.6. From all three algorithms, k-NNs and RFs performed better than SVR although in some occasions, k-NNs performed better than RFs for data of certain participants and vice versa. It is clearly shown that our dataset got overall better results, which means that the signal processing pipeline was more efficient. More analytically, results are shown in tables B.1, B.2 and B.3 for the DEAP dataset and for our data in B.4, B.5 and B.6 of the appendix.

In section 3.2.2, the VAE is utilised as well for the denoising of EEG signals, more specifically determining the latent representation of signals with the least noise. The machine learning results for this preprocessing approach on the DEAP dataset are given in figures 4.7, 4.8 and 4.9 and for our data in 4.10, 4.11 and 4.12. From the results of the algorithm, almost the

same observations are encountered as with ICA. There is this slight conflict between k-NNs and RFs in which performs better than the other. But now, SVR achieves slightly better results than both. Furthermore, again our dataset got better results, which means that the signal processing pipeline was more efficient. More analytically, results are shown in tables B.7, B.8 and B.9 for the DEAP dataset and for our data in B.10, B.11 and B.12 of the appendix. While training the network, we were plotting the loss function for all 50 epochs of each participant as shown in figure 4.13. This is the loss function from training of a random participant. Loss function for DEAP in appendix C.

Overall, for both datasets in SVR the VAE has performed better although in k-NNs results were very identical for the two datasets but ICA was slightly better for our data. For the DEAP dataset, sometimes ICA was better than VAE for data of certain participants and vice versa. Same observations were encountered for RFs on the DEAP dataset, although for our data ICA was more efficient. Hyperparameter tuning through 5-fold cross validation was done and the mean squared error (MSE) was plotted against several values of k for k-NNs, and against the number of decision trees, criteria and maximum depth for RFs. An illustration in figures 4.14, 4.15, 4.16 and 4.17 is shown. The rest of the curves for all other models in appendix D.

A brief analysis of results for pseudo labelling is done after performing signal decoding with ICA and VAE. Overall, the classifier worked really well in predicting the correct label. Correct labels were assessed based on the method described in section 3.4. Although, it did make some mistakes and those were only for specific cases. For ICA, the pseudo classifier mainly did mistakes in predicting emotions with negative valence or neutral dominance. In predicting arousal, it done almost all predictions correct except from a few discrepancies. For VAE, the classifier made similar mistakes as in ICA. Again, mainly on predicting emotions with negative valence or neutral dominance. The rest were predicted correctly and can be re-used effectively in the regression modelling.

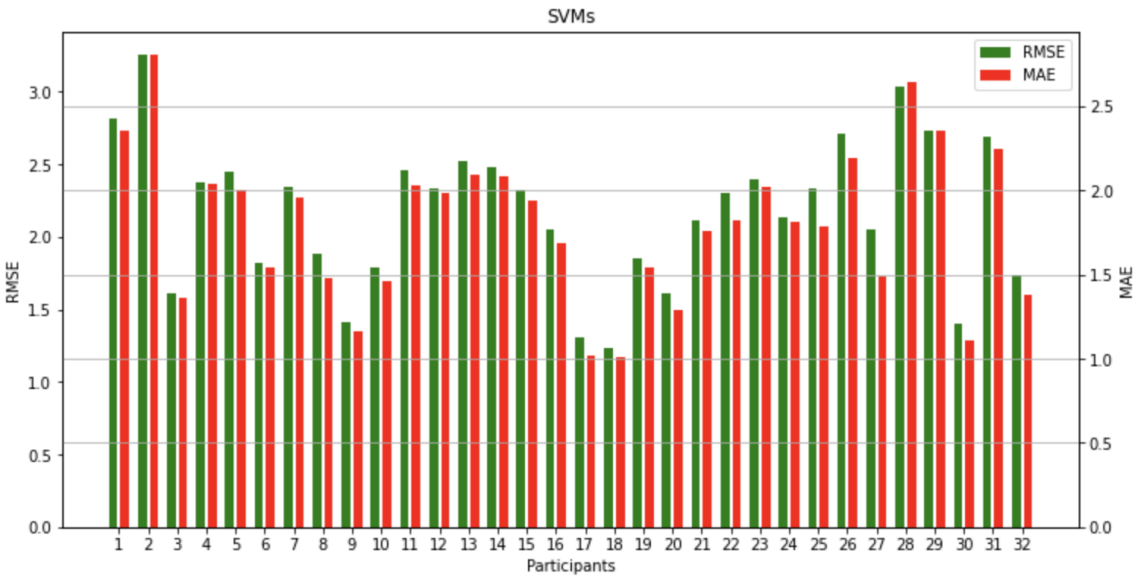

Figure 4.1: SVR results' barplots with ICA for the DEAP dataset. It is shown that SVR achieves the least good results from all three algorithms.

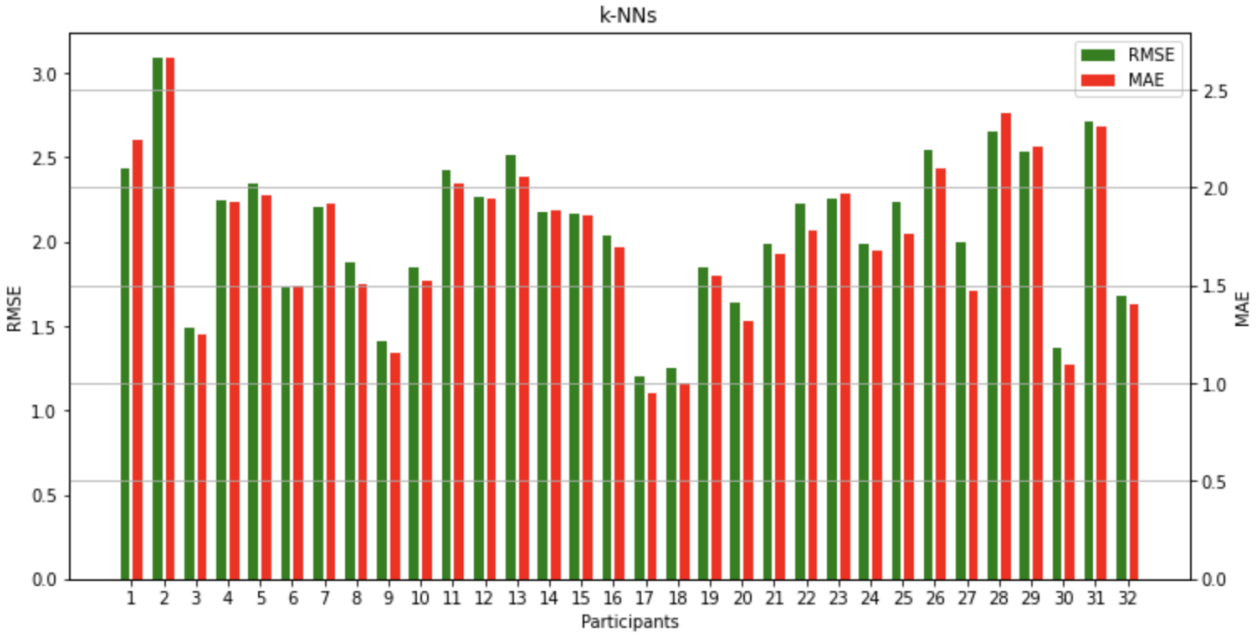

Figure 4.2: k-NNs results' barplots with ICA for the DEAP dataset. It almost achieves better results from SVR and almost similar with RFs.

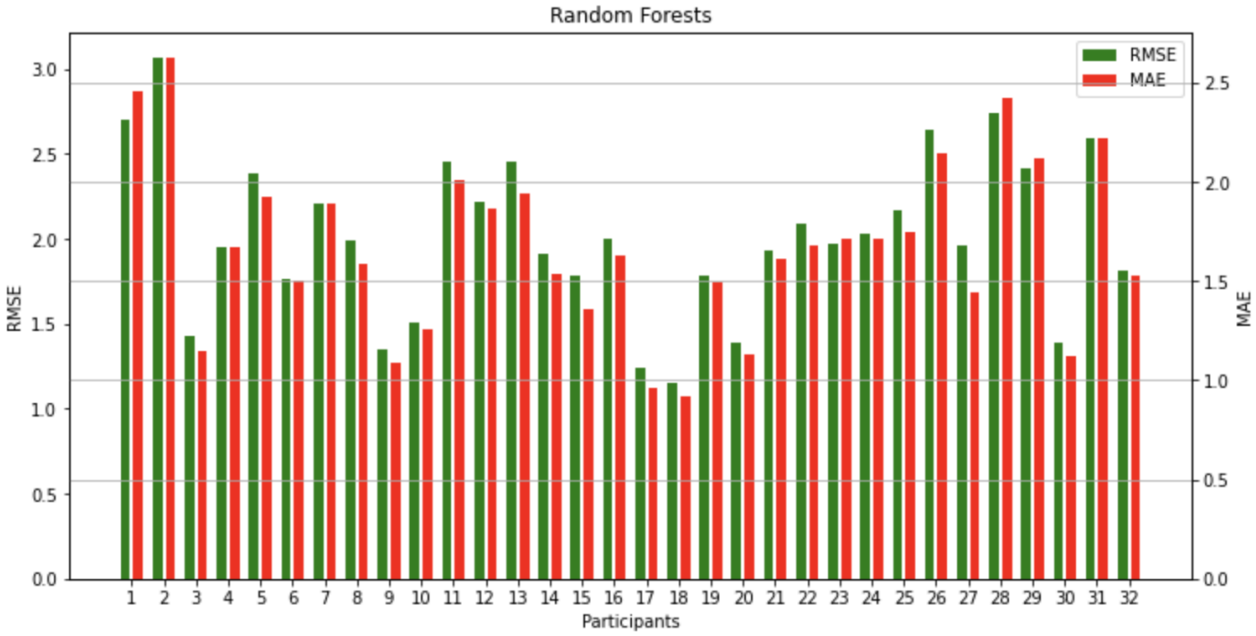

Figure 4.3: RFs results' barplots with ICA for the DEAP dataset. It almost achieves better results from SVR and almost similar with k-NNs.

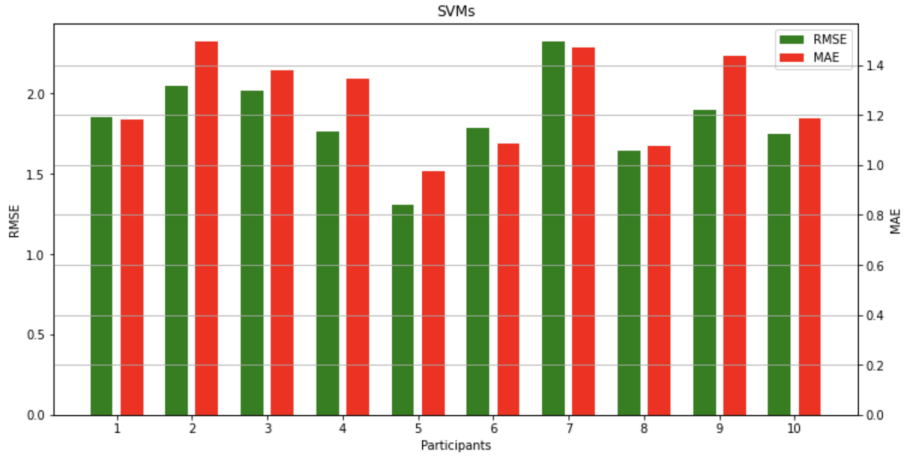

Figure 4.4: SVR results' barplots with ICA for our dataset. Overall results for our dataset outperformed those of DEAP. It is shown that SVR achieves the least good results from all three algorithms.

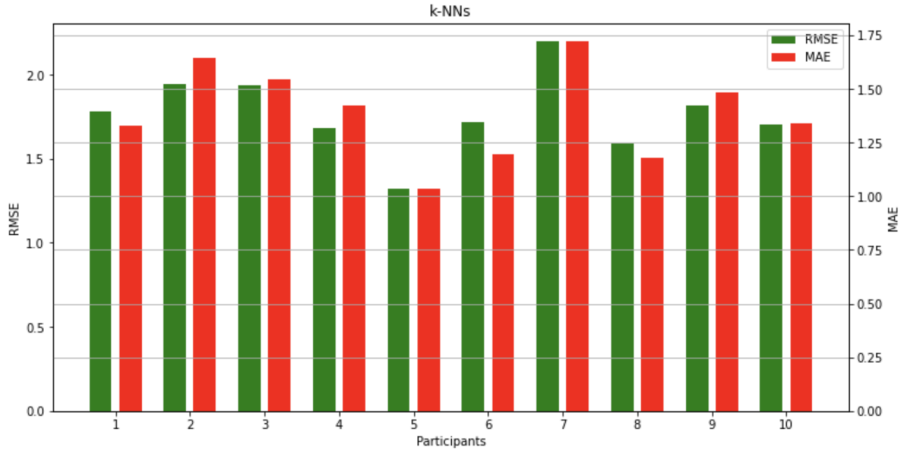

Figure 4.5: k-NNs results' barplots with ICA for our dataset. Overall results for our dataset outperformed those of DEAP. It almost achieves better results from SVR and almost similar with RFs.

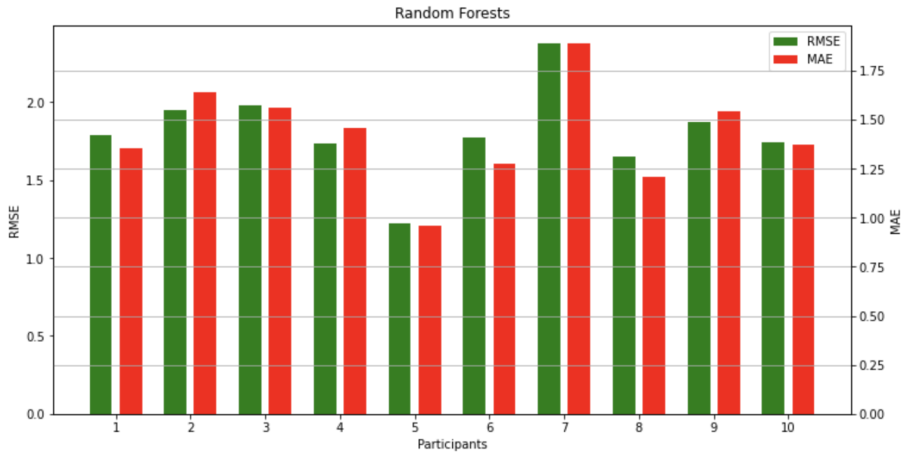

Figure 4.6: RFs results' barplots with ICA for our dataset. Overall results for our dataset outperformed those of DEAP. It almost achieves better results from SVR and almost similar with k-NNs.

Figure 4.7: SVR results' barplots with VAE for the DEAP dataset. It is shown that SVR achieves the best results from all three algorithms.

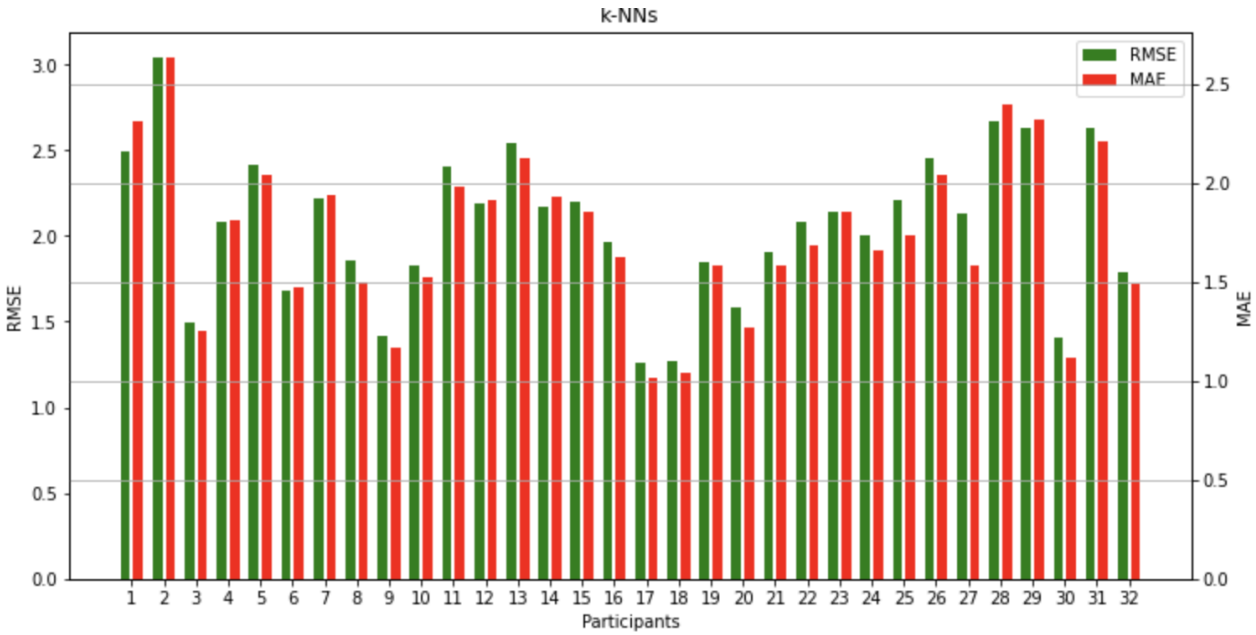

Figure 4.8: k-NNs results' barplots with VAE for the DEAP dataset. SVR achieves slightly better results from k-NNs, although k-NNs achieves almost similar with RFs.

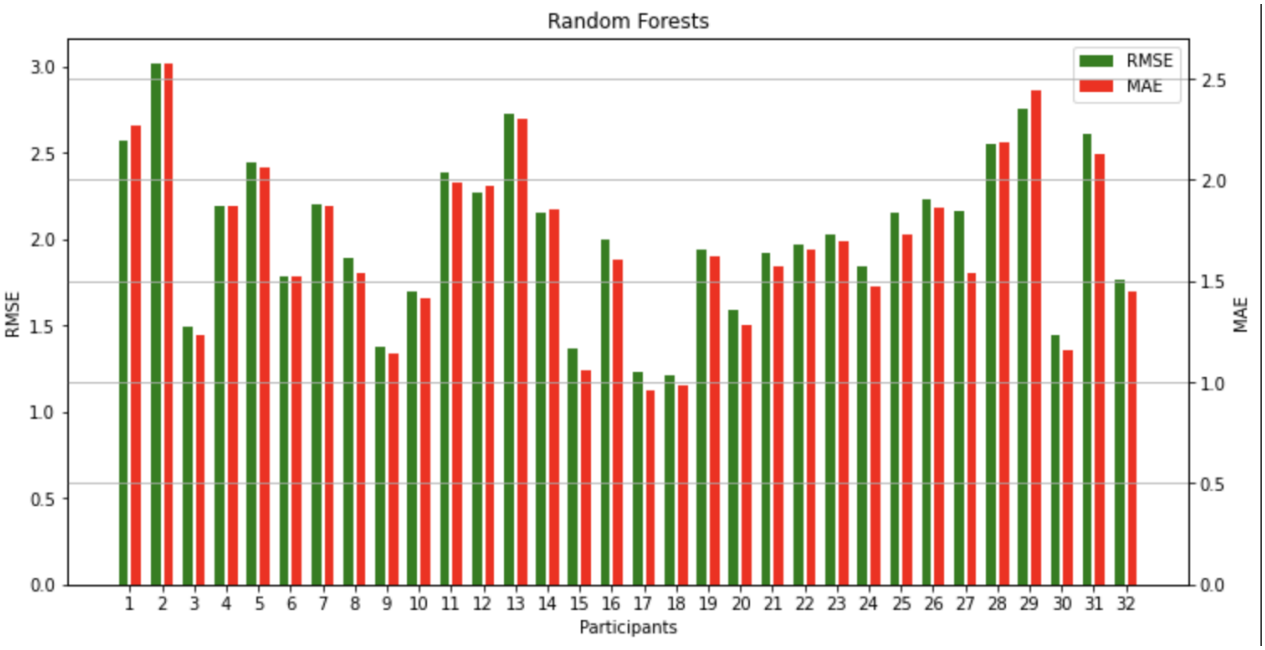

Figure 4.9: RFs results' barplots with VAE for the DEAP dataset. SVR achieves slightly better results from RFs, although RFs achieves almost similar with k-NNs.

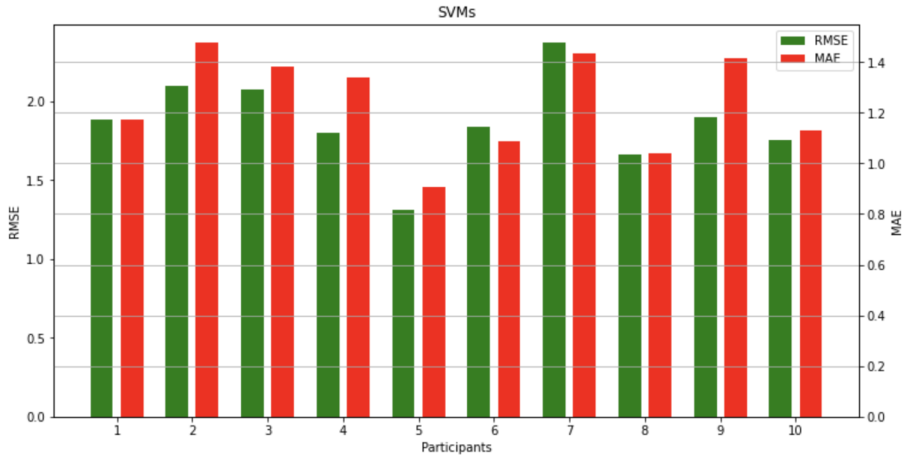

Figure 4.10: SVR results' barplots with VAE for our dataset. Overall results for our dataset outperformed those of DEAP. It is shown that SVR achieves the best results from all three algorithms.

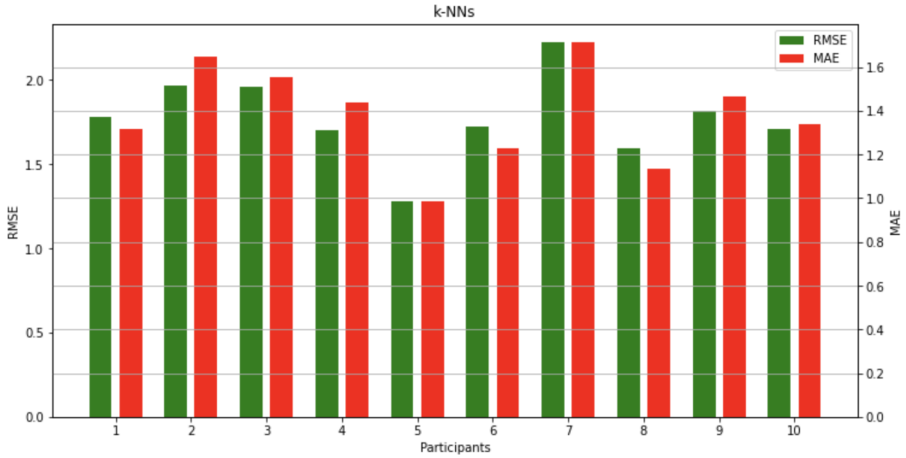

Figure 4.11: k-NNs results' barplots with VAE for our dataset. Overall results for our dataset outperformed those of DEAP. SVR achieves slightly better results from k-NNs, although k-NNs achieves almost similar with RFs.

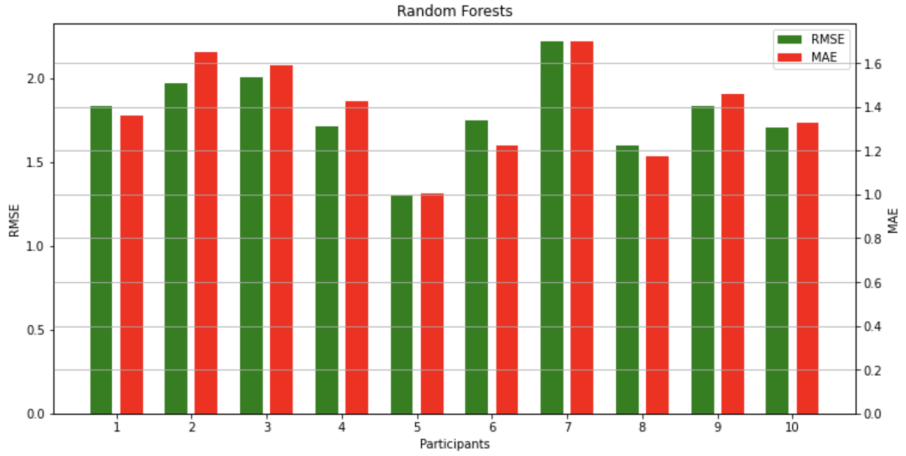

Figure 4.12: RFs results' barplots with VAE for our dataset. Overall results for our dataset outperformed those of DEAP. SVR achieves slightly better results from RFs, although RFs achieves almost similar with k-NNs.

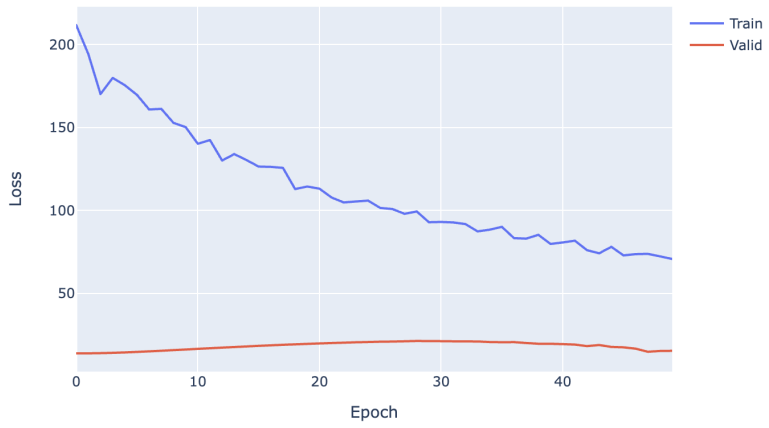

Figure 4.13: Loss function visualisation for VAE for our dataset. It clearly indicates that throughout epochs the validation loss is smaller than the training loss, meaning that it's not overfitting and both losses are small meaning that it's not underfitting.

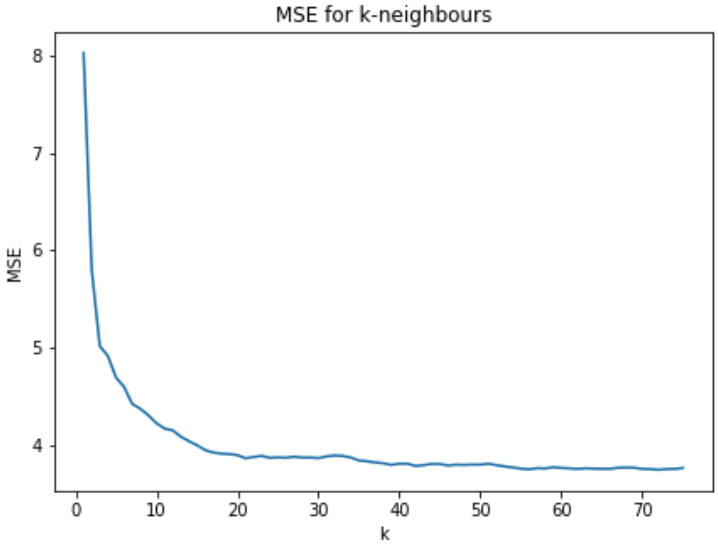

Figure 4.14: Tuning for k neighbours in k-NNs with ICA on our dataset. Computing the MSE for separate models while increasing k we observe that the optimal k is 71.

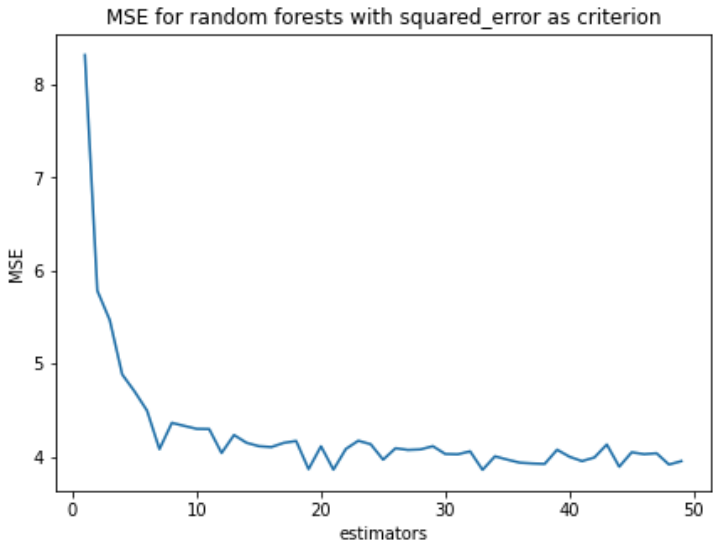

Figure 4.15: Tuning of best criterium for RFs using ICA on our dataset. Computing the MSE for separate models while changing criteria we find that the optimal criterium is 'squared error'.

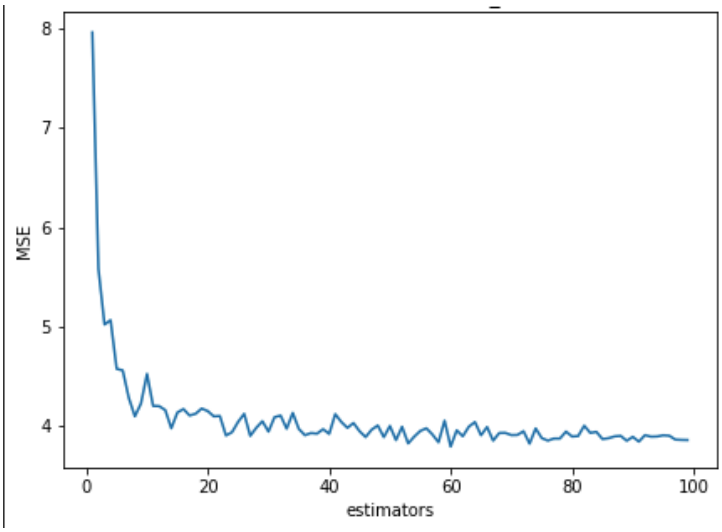

Figure 4.16: Optimal number of decision trees for RFs using ICA on our dataset. Computing the MSE for separate models while increasing decision trees we observe that the optimal number is 59.

Figure 4.17: Optimal maximum depth of decision trees for RFs using ICA on our dataset. Computing the MSE for separate models while increasing the depth we observe that the optimal maximum depth is 1.

Chapter 5

Discussion

It is fascinating how such good results have been achieved for our dataset. The reason is because, as mentioned in chapter 4 many issues were encountered during physical data collection of EEG data, general wear-to-tear of the EPOC device has deteriorated the EEG signal recordings leading in final data to be recorded considered as 'unreliable'. In other words, data for each component of the device has generated noise instead. A clear indication is also shown in figures 3.7 and 3.8 where the top right heatmaps appear to be too noisy from the red spots that exist. This effect has occured in all 14 components that we collected data from, and it is a clear proof that overall data that were collected are unreliable. Although, during preprocessing we only retained frequency bands between 4-50 Hz and filtered out the rest as it should have been done whereas for the DEAP dataset, EEG data has been already preprocessed before hand but retained frequencies between 4-45 Hz. That could be considered a case why results are not as good as those from our data, because data between 45-50 Hz mainly consist the high γ band that contains essential information for emotions within EEG decoding.

Regarding **(H1)** stated in section 2.5, the effectiveness of a VAE on EEG signal decoding with respect to ICA has been disproved. More analytically, it has only proven right when observing machine learning results with SVR mentioned in section 4.2. This observation may depend on how hyperparameters are tuned. With ICA, k-NNs and RFs are better than SVR. ICA itself still allows data to include large amounts of noise, hence the regularisation parameter (1.0) used to prevent overfitting while training the model was too small. One the other hand, VAE with ICA extracts much more information from the raw signal meaning that the regularization parameter used (1.0) was appropriate enough to explain why it slightly achieves better results than the other two algorithms. But in general, the aim was to out-beat results that undertook only ICA as part of the preprocessing, including all possible cases of hyperparameters. Although, several reasons might be the case for that. For example, inducing and comparing more feature extraction methods such as fractal dimensions and differential entropy [37] could have improved results.

As it wasn't one of the points of investigation, PSD has only been chosen as it is also considered one of the most essential frequency domain feature extraction methods. Furthermore, regarding normalisation of the EEG data for ICA it was performed right before regression modelling and right after the preprocessing and filtering of the signals. Although, for VAE normalisation was conducted before preprocessing as it is essential to do that only then if deep

learning is implemented. Or that the inclusion of Dropout layers might have excluded important information. But, the ratio between non-trainable paramaters with respect trainable ones may be considered negligible as well. Another factor could be that other types of autoencoders such as denoising autoencoders, LSTM-AE, RBM could achieved better results. Although, the VAE was preferred and used due to the inspiration given from the work conducted by Li [22] *et al.*. And it was one of the first pieces of work that utilised a VAE for such a purpose. Hence, this domain is still under investigation and in the near future it might be considered one of the most robust techniques.

Additionally, the VAE may not have performed really well as it may required more training data and overall in deep learning it is essential. One utility of the VAE is that its decoder can be used to generate synthetic data out of the compressed inputs, hence aid in the desire of data augmentation. With the main assumption of VAEs that the latent vectors follow a Gaussian distribution, during their decoding the synthetic data is constrained to follow a Gaussian distribution as well alike the original data given as input. Hence, in theory adding some Gaussian noise to the latent vectors can be perfectly represented as normal EEG data and with more data the network can be trained more effectively.

For **(H2)** in 2.5, results have shown that it has been both proved and disproved. It certainly means, that the pseudo labels turnt to be correct with the respect to the ANEW and efficiently were used to augment the target variables for further regression modelling. More specifically, mis-classifications of pseudo labels were mainly done in emotions of negative valence and neutral dominance although in arousal almost all pseudo labels were predicted correctly except from a few discrepancies. That can be justified from the fact that there were more training examples for arousal, as participant had the fewest misclassifications as mentioned in section 4.1. There hasn't been much of an improvement between results after implementing ICA and VAE for signal processing. In both signal processing pipelines, the classifier predicted similar labels and did also mistakes in certain cases as those stated in section 4.2. These may be from the issues stated on the above paragraph as it again depends from the signal processing pipelines.

But if we pay particular attention for the emotion of 'embarrassment' (based on the ANEW [4] 'embarrassement' has valence approximately to be 3) as shown in table 4.2, participants made the most mistakes in predicting valence which means that it would have been hard for the classifier to make those predictions correctly as the amount of unlabelled data is larger than labelled data which this may give an explanation of the classifier making those mistakes. The same applies for neutral dominance. Based on table 4.4 most mis-classifications from participants were done on 'shame' but considering 'anger' and 'disgust' (based on the ANEW [4] 'anger' and 'disgust' have dominance scores approximately to be 6 and 5 respectively) which still had a lot of mis-classifications and combining these two might be the issue for the classifier to predict incorreclty emotions of neutral dominance. This is explained exactly as in for the case of negative valence right above. As a matter of fact, a semi-supervised learning approach makes the problem formulation more correct and it also proves that for issues encountered in data collection it can be considered a solution. Whereas most of the researchers use it because data labelling in most cases is expensive and time consuming, we use it as we believe that target data collected by participants can be considered unreliable and the usage of the ANEW for sanity-checking proved that an approach like pseudo labelling is essential.

The algorithm of ICA itself may have been the leader in most of the observations detected

within the results. It is one of the standard approaches that is utilised in the field of EEG emotion recognition and more certainly it is used to remove huge sources of artifacts such as the EOG, EMG and ECG. The resultant signal will still contain other types of artifacts and in most cases there's not much to do to remove those as well. On the other hand, the idea of utilising a VAE in addition of implementing ICA to remove the major artifacts was to determine and remove the sources of data that contain artifacts that isn't that easy to remove in general. We could have obtained the better results if we were to follow any alternatives as discussed above.

Overall, in terms of the regression metrics used results for both signal processing pipelines and for all machine learning algorithm are really good which means that most of the work has been done correctly and only a few discrepancies may have got us better results. In general, data from every participant are very different and that might explain the similarities as well as the inconsistencies of the results that were obtained. Hence, it may have required to conduct separate hyperparameter tuning for different data among participants. That would have explained results much better and the reason this was not done is due to the limited amounts of time that we had for running the codes and obtaining the results.

Chapter 6

Conclusions

In this investigation, we depicted that emotion recognition through EEG decoding is possible, more certainly applying regression modelling to all three dimensions of valence, arousal and dominance for getting more accurate predictions and in generalising the problem formulation making it even broader.

Unsupervised learning methods such as ICA and with the implementation of deep learning such as VAE are considered the most crucial tasks for the emotion recognition pipeline. To be able to learn from EEG data which are considered very noisy, it must be ensured that data preprocessing is done correctly. Otherwise, the regression models would struggle in predicting the accurate target variables. The presence of noisy labels allowed to view this investigation as more of a real life scenario and what types of issues a researcher may encounter if pursuing physical EEG data collection.

Going more deep and investigating more the human brain structure, one very interesting domain would be the decoding of mental illnesses such as Alzheimer's disease or depression. Inspired by Li [22] *et al.*, they state that it certainly is possible with the aid of variational autoencoders for decoding EEG. As a next task, we want to investigate how to construct a VAE to decode EEG this way and detect which information from an EEG is essential for this task.

Bibliography

[1] Alchalabi, A.E., Elsharnouby, M., Shirmohammadi, S. and Eddin, A.N., 2017. Feasibility of detecting adhd patients' attention levels by classifying their eeg signals. *2017 ieee international symposium on medical measurements and applications (memea)*. IEEE, pp.314–319.

[2] Awad, M. and Khanna, R., 2015. Support vector regression. *Efficient learning machines*. Springer, pp.67–80.

[3] Birla, D., 2019. Basics of autoencoders. `https://medium.com/@birla.deepak26/autoencoders-76bb49ae6a8f`. [Online], [Accessed 8th May].

[4] Bradley, M.M. and Lang, P.J., 1999. *Affective norms for english words (anew): Instruction manual and affective ratings*. Technical report C-1, the center for research in psychophysiology ....

[5] Cook, I.A., Warren, C., Pajot, S.K., Schairer, D. and Leuchter, A.F., 2011. Regional brain activation with advertising images. *Journal of neuroscience, psychology, and economics*, 4(3), p.147.

[6] Du, Z., Wu, S., Huang, D., Li, W. and Wang, Y., 2019. Spatio-temporal encoder-decoder fully convolutional network for video-based dimensional emotion recognition. *Ieee transactions on affective computing*, 12(3), pp.565–578.

[7] Fayek, H.M., Lech, M. and Cavedon, L., 2016. Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels. *2016 international joint conference on neural networks (ijcnn)*. IEEE, pp.566–570.

[8] Ferretti, V. and Papaleo, F., 2019. Understanding others: Emotion recognition in humans and other animals. *Genes, brain and behavior*, 18(1), p.e12544.

[9] Galvão, F., Alarcão, S.M. and Fonseca, M.J., 2021. Predicting exact valence and arousal values from eeg. *Sensors*, 21(10), p.3414.

[10] Gašpar, T., Labor, M., Jurić, I., Dumančić, D., Ilakovac, V. and Heffer, M., 2011. Comparison of emotion recognition from facial expression and music. *Collegium antropologicum*, 35(1), pp.163–167.

[11] Gunes, H. and Schuller, B., 2013. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and vision computing*, 31(2), pp.120–136.

[12] Hadjidimitriou, S.K. and Hadjileontiadis, L.J., 2013. Eeg-based classification of music

appraisal responses using time-frequency analysis and familiarity ratings. *Ieee transactions on affective computing*, 4(2), pp.161–172.

[13] Han, B.j., Rho, S., Dannenberg, R.B. and Hwang, E., 2009. Smers: Music emotion recognition using support vector regression. *Ismir*. Citeseer, pp.651–656.

[14] Islam, M., Ahmed, T., Mostafa, S.S., Yusuf, M.S.U. and Ahmad, M., 2013. Human emotion recognition using frequency and statistical measures of eeg signal [Online]. *2013 international conference on informatics, electronics and vision (iciev)*. pp.1–6. Available from: `https://doi.org/10.1109/ICIEV.2013.6572658`.

[15] James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. *An introduction to statistical learning: with applications in r*. Spinger.

[16] Kim, Y.E., Schmidt, E.M., Migneco, R., Morton, B.G., Richardson, P., Scott, J., Speck, J.A. and Turnbull, D., 2010. Music emotion recognition: A state of the art review. *Proc. ismir*. vol. 86, pp.937–952.

[17] Koelstra, S., Muhl, C., Soleymani, M., Lee, J.S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A. and Patras, I., 2011. Deap: A database for emotion analysis; using physiological signals. *Ieee transactions on affective computing*, 3(1), pp.18–31.

[18] Kosti, R., Alvarez, J.M., Recasens, A. and Lapedriza, A., 2017. Emotion recognition in context. *Proceedings of the ieee conference on computer vision and pattern recognition*. pp.1667–1675.

[19] Lee, D., 2013. The simple and efficient semi-supervised learning method for deep neural networks. *In workshop on challenges in representation learning*, 3(2), p.896.

[20] Li, C., Hou, Y., Song, R., Cheng, J., Liu, Y. and Chen, X., 2022. Multi-channel eeg-based emotion recognition in the presence of noisy labels. *Science china information sciences*, 65(4), pp.1–16.

[21] Li, X., Zhao, Z., Song, D., Zhang, Y., Niu, C., Zhang, J., Huo, J. and Li, J., 2019. Variational autoencoder based latent factor decoding of multichannel eeg for emotion recognition. *2019 ieee international conference on bioinformatics and biomedicine (bibm)*. IEEE, pp.684–687.

[22] Li, X., Zhao, Z., Song, D., Zhang, Y., Pan, J., Wu, L., Huo, J., Niu, C. and Wang, D., 2020. Latent factor decoding of multi-channel eeg for emotion recognition through autoencoder-like neural networks. *Frontiers in neuroscience*, 14, p.87.

[23] Liu, Y. and Sourina, O., 2014. Real-time subject-dependent eeg-based emotion recognition algorithm. *Transactions on computational science xxiii*. Springer, pp.199–223.

[24] Liu, Y. and Wang, X., 2020. Differences in driving intention transitions caused by driver's emotion evolutions. *International journal of environmental research and public health* [Online], 17(19). Available from: `https://doi.org/10.3390/ijerph17196962`.

[25] Martínez, F., Frías, M.P., Pérez, M.D. and Rivera, A.J., 2019. A methodology for applying k-nearest neighbor to time series forecasting. *Artificial intelligence review*, 52(3).

[26] Monakhova, Y.B. and Rutledge, D.N., 2020. Independent components analysis (ica) at the "cocktail-party" in analytical chemistry. *Talanta*, 208, p.120451.

[27] Office, I.P., 2012. Exceptions to copyright:education and teaching. `https://www.eecs.qmul.ac.uk/mmv/datasets/deap/readme.html`. [Online], [Accessed 3rd September 2022].

[28] Olszanowski, M., Pochwatko, G., Kuklinski, K., Scibor-Rylski, M., Lewinski, P. and Ohme, R.K., 2015. Warsaw set of emotional facial expression pictures: a validation study of facial display photographs. *Frontiers in psychology*, 5, p.1516.

[29] Pontifex, M.B., Gwizdala, K.L., Parks, A.C., Billinger, M. and Brunner, C., 2017. Variability of ica decomposition may impact eeg signals when used to remove eyeblink artifacts. *Psychophysiology*, 54(3), pp.386–398.

[30] Rahman, M.M., Sarkar, A.K., Hossain, M.A., Hossain, M.S., Islam, M.R., Hossain, M.B., Quinn, J.M. and Moni, M.A., 2021. Recognition of human emotions using eeg signals: A review. *Computers in biology and medicine*, 136, p.104696.

[31] Reyes, B.N., Segal, S.C. and Moulson, M.C., 2018. An investigation of the effect of race-based social categorization on adults' recognition of emotion. *Plos one*, 13(2), p.e0192418.

[32] Shaw, L. and Bagha, S., 2012. Online emg signal analysis for diagnosis of neuromuscular diseases by using pca and pnn. *International journal of engineering science and technology*, 4(10), pp.4453–4459.

[33] Shih, J.J., Krusienski, D.J. and Wolpaw, J.R., 2012. Brain-computer interfaces in medicine. *Mayo clinic proceedings*. Elsevier, vol. 87, pp.268–279.

[34] Solomon Jr, O.M., 1991. *Psd computations using welch's method.[power spectral density (psd)]*. Sandia National Labs., Albuquerque, NM (United States).

[35] Sutton, T.M., Herbert, A.M. and Clark, D.Q., 2019. Valence, arousal, and dominance ratings for facial stimuli. *Quarterly journal of experimental psychology*, 72(8), pp.2046–2055.

[36] Tandle, A.L., Joshi, M.S., Dharmadhikari, A.S. and Jaiswal, S.V., 2018. Mental state and emotion detection from musically stimulated eeg. *Brain informatics*, 5(2), pp.1–13.

[37] Thammasan, N., Fukui, K.i. and Numao, M., 2016. Application of deep belief networks in eeg-based dynamic music-emotion recognition [Online]. *2016 international joint conference on neural networks (ijcnn)*. pp.881–888. Available from: `https://doi.org/10.1109/IJCNN.2016.7727292`.

[38] Thammasan, N., Moriyama, K., Fukui, K.i. and Numao, M., 2017. Familiarity effects in eeg-based emotion recognition. *Brain informatics*, 4(1), pp.39–50.

[39] Tharwat, A., 2020. Independent component analysis: An introduction. *Applied computing and informatics*.

[40] Toisoul, A., Kossaifi, J., Bulat, A., Tzimiropoulos, G. and Pantic, M., 2021. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature machine intelligence*, 3(1), pp.42–50.

[41] Tracy, J.L., Robins, R.W. and Schriber, R.A., 2009. Development of a facs-verified set of basic and self-conscious emotion expressions. *Emotion*, 9(4), p.554.

[42] Wei Li, Shengchen Li, X.S.Z.L., 2019. *Proceedings of the 6th conference on sound and music technology (csmt)*, vol. 568. Available from: `https://link.springer.com/content/pdf/10.1007/978-981-13-8707-4.pdf`.

[43] Zhang, G. and Etemad, A., 2021. Deep recurrent semi-supervised eeg representation learning for emotion recognition [Online]. *2021 9th international conference on affective computing and intelligent interaction (acii)*. pp.1–8. Available from: `https://doi.org/10.1109/ACII52823.2021.9597449`.

Appendix A

Data acquisition procedure

Figure A.1: Step 1: Participant watching a single image for 4 seconds without any movements.

Figure A.2: Step 2: Participant completing the self-assessment manikin scores of a single image for 10 seconds with pen and paper.

Figure A.3: Step 3: Participant stabilising for 5 seconds to prepare for the appearance of the next image.

Appendix B

Results Tables

Table B.1: SVR results' barplots with ICA for the DEAP dataset. It is shown that SVR achieves the least good results from all three algorithms.

	RMSE	MAE
Participant 1	2.8125887388133406	2.348423674728773
Participant 2	3.2534742486328274	2.8023355363078846
Participant 3	1.6069437406139784	1.361382092167042
Participant 4	2.3776548350427382	2.040767647469765
Participant 5	2.451550682552658	1.9994665611655897
Participant 6	1.8251216484915451	1.5440365528928541
Participant 7	2.34723547156154	1.9602887654665704
Participant 8	1.8795214034719814	1.4746762462035334
Participant 9	1.4092723175033184	1.1625537053647197
Participant 10	1.7931116490045955	1.4561952563397362
Participant 11	2.4551254865951173	2.025395687062128
Participant 12	2.335657603498315	1.9812268741807941
Participant 13	2.5232249413964705	2.094655030179127
Participant 14	2.4822706805022854	2.079838291431811
Participant 15	2.323887256006318	1.9358643543709342
Participant 16	2.04563185354725	1.684060264128996
Participant 17	1.305109224882871	1.0175257548933536
Participant 18	1.2331320558420695	1.0106645943320371
Participant 19	1.8505309754666033	1.5403217790960795
Participant 20	1.6074698953004334	1.2866803424006406
Participant 21	2.111809214161405	1.7599487653077601
Participant 22	2.2974356338297612	1.82119794084753
Participant 23	2.3964791028039594	2.019673532104834
Participant 24	2.1338539513537134	1.8139995061136138
Participant 25	2.329774517052737	1.7873261314105315
Participant 26	2.70547192479734	2.1909332966962736
Participant 27	2.046132798357023	1.4862723329847798
Participant 28	3.036846110035317	2.6399593080920516
Participant 29	2.7354558353345753	2.351229588670267
Participant 30	1.3964277470477673	1.1051704580112534
Participant 31	2.685211620922858	2.2489744285437543
Participant 32	1.7336361168342702	1.378984436764571

Table B.2: k-NNs results' barplots with ICA for the DEAP dataset. It almost achieves better results from SVR and almost similar with RFs.

	RMSE	MAE
Participant 1	2.437653909403722	2.242677083333333
Participant 2	3.0891316681052343	2.6620156250000004
Participant 3	1.4860739192110783	1.2517864583333334
Participant 4	2.2452006539644738	1.9254583333333335
Participant 5	2.3484454562961252	1.9585625000000002
Participant 6	1.7258062927298767	1.501536458333333
Participant 7	2.2088530154539954	1.9136822916666667
Participant 8	1.8750029852406793	1.5063593750000004
Participant 9	1.4054575695543812	1.1553281249999998
Participant 10	1.8438911404311311	1.5197187500000002
Participant 11	2.424025727826886	2.0241562500000003
Participant 12	2.2661334975340344	1.9395833333333332
Participant 13	2.5123934102338468	2.055536458333333
Participant 14	2.1784715709452502	1.8866562500000004
Participant 15	2.168333792465283	1.8599218750000002
Participant 16	2.040478477064775	1.6971770833333337
Participant 17	1.199367963708299	0.9467760416666667
Participant 18	1.2486889139838833	1.0020624999999999
Participant 19	1.8480011724383003	1.5503749999999998
Participant 20	1.6413313112553758	1.3201718750000002
Participant 21	1.99149898478011	1.6592760416666668
Participant 22	2.220965194793555	1.7802968750000001
Participant 23	2.251988314899238	1.9656093750000003
Participant 24	1.987264005086254	1.6806822916666662
Participant 25	2.2313843527185746	1.7610677083333333
Participant 26	2.543516868411891	2.098166666666667
Participant 27	1.9977913431214585	1.4765572916666672
Participant 28	2.6511840373935205	2.384828125
Participant 29	2.5356450995680646	2.2065052083333327
Participant 30	1.3685231725579676	1.0939322916666665
Participant 31	2.7084365264795762	2.3118020833333333
Participant 32	1.683581010682661	1.4024166666666666

Table B.3: RFs results' barplots with ICA for the DEAP dataset. It almost achieves better results from SVR and almost similar with k-NNs.

	RMSE	MAE
Participant 1	2.701540830940607	2.4551568627450977
Participant 2	3.0648441171691676	2.6240277777777776
Participant 3	1.42456241316831	1.145277777777778
Participant 4	1.948006843562332	1.6698186274509805
Participant 5	2.381059801900985	1.9227818627450979
Participant 6	1.7646225169806036	1.5034861111111109
Participant 7	2.208993094139451	1.8913848039215686
Participant 8	1.9872007082557992	1.5905629084967323
Participant 9	1.3473531327772672	1.085359477124183
Participant 10	1.5067564908734628	1.255468137254902
Participant 11	2.4528725340507718	2.0127385620915033
Participant 12	2.2193953686027745	1.8682426470588236
Participant 13	2.4531873324299966	1.9408700980392155
Participant 14	1.909100798051863	1.5362246732026146
Participant 15	1.7827875000079352	1.3620457516339872
Participant 16	1.998286221018018	1.6299599673202614
Participant 17	1.2428087654708098	0.963845588235294
Participant 18	1.1534690882464291	0.9189665032679738
Participant 19	1.787173434158715	1.4930882352941177
Participant 20	1.3851668899351353	1.1319746732026144
Participant 21	1.9290151068582915	1.6088055555555556
Participant 22	2.087814171823618	1.6761756535947712
Participant 23	1.9749330242586896	1.7108088235294119
Participant 24	2.025847039905673	1.7095972222222227
Participant 25	2.1699118863850497	1.7471176470588234
Participant 26	2.6415961527767613	2.1454068627450984
Participant 27	1.9637388222156336	1.439152777777778
Participant 28	2.7394253604994954	2.4208954248366013
Participant 29	2.4171323085950966	2.1183096405228756
Participant 30	1.3873976981981506	1.123856209150327
Participant 31	2.594684773531069	2.2173521241830065
Participant 32	1.8145635629107604	1.5247434640522872

Table B.4: SVR results' barplots with ICA for our dataset. Overall results for our dataset outperformed those of DEAP. It is shown that SVR achieves the least good results from all three algorithms.

	RMSE	MAE
Participant 1	1.8497471272201338	1.1838935146724585
Participant 2	2.0475476853039973	1.493909513047164
Participant 3	2.015656563384656	1.3781311185096712
Participant 4	1.7638692738749346	1.3467723189890313
Participant 5	1.3051253743249482	0.9731570712506441
Participant 6	1.7850131812516838	1.0847840957495982
Participant 7	2.3213821692452363	1.4698776510224096
Participant 8	1.642484380578131	1.0750131102646938
Participant 9	1.8975779816248204	1.4354297703148997
Participant 10	1.7453691619747966	1.185103864933652

Table B.5: k-NNs results' barplots with ICA for our dataset. Overall results for our dataset outperformed those of DEAP. It almost achieves better results from SVR and almost similar with RFs.

	RMSE	MAE
Participant 1	1.7787606162850602	1.326490860053941
Participant 2	1.9475355591729961	1.644329132690882
Participant 3	1.937096196747662	1.542525327403015
Participant 4	1.6844108453731255	1.4222881146528294
Participant 5	1.3183973432373621	1.0350877192982457
Participant 6	1.7191301456447945	1.1956016802569804
Participant 7	2.197685470942759	1.720286632073141
Participant 8	1.5896720036907024	1.1772176921176178
Participant 9	1.8198056330846264	1.4835186557944153
Participant 10	1.702473946230963	1.3372374598468

Table B.6: RFs results' barplots with ICA for our dataset. Overall results for our dataset outperformed those of DEAP. It almost achieves better results from SVR and almost similar with k-NNs.

	RMSE	MAE
Participant 1	1.7998598857909203	1.3629791816285988
Participant 2	1.9689902748106065	1.6561941857823623
Participant 3	1.9936401544063846	1.585147968769272
Participant 4	1.7424343760430745	1.4739986821611792
Participant 5	1.2159628123228945	0.9515610361094184
Participant 6	1.7549556563244912	1.2506746758193674
Participant 7	2.3838676833248913	1.8997006715606464
Participant 8	1.6279448618597305	1.2089926183155786
Participant 9	1.888816219724183	1.5269345183822498
Participant 10	1.7421717247116313	1.3625056008579577

Table B.7: SVR results' barplots with VAE for the DEAP dataset. It is shown that SVR achieves the best good results from all three algorithms.

	RMSE	MAE
Participant 1	2.8038680877900926	2.4214263404915806
Participant 2	3.0998521544869027	2.6280479184118923
Participant 3	1.526861814880475	1.263896760900124
Participant 4	2.2382280476088123	1.9262900430468477
Participant 5	2.4658687796487433	2.083293588395687
Participant 6	1.7443378975588604	1.4810052529703537
Participant 7	2.271469635906833	1.9034568473172973
Participant 8	1.9038098086182778	1.5485902170240962
Participant 9	1.3815233914746603	1.1327557238420662
Participant 10	1.8802638038127042	1.585237617121716
Participant 11	2.3701521461776496	1.9668890680276334
Participant 12	2.3484819462798963	2.0356257485265234
Participant 13	2.5852083240325565	2.156826549224535
Participant 14	2.3677828561868566	2.0251090022260385
Participant 15	2.3293911880656784	1.9694179921380413
Participant 16	1.9730440316359767	1.627759419805036
Participant 17	1.3043561504641266	1.0442297709937287
Participant 18	1.2383402833289001	1.0190318803460243
Participant 19	1.8767801643432458	1.56656010602779
Participant 20	1.5525651289981726	1.2576513212865101
Participant 21	1.8380254866060297	1.4769381053377177
Participant 22	2.141639125452712	1.7060436250565028
Participant 23	2.1233925464616883	1.7739607937085597
Participant 24	1.996802140085329	1.6804241280497667
Participant 25	2.237704214521326	1.763320747367189
Participant 26	2.4302314275188106	2.003362343941709
Participant 27	2.0905321676442745	1.5238192011527998
Participant 28	2.9163576714594712	2.534780756875604
Participant 29	2.812561404816083	2.446195320604543
Participant 30	1.4414599112830992	1.1374959211742364
Participant 31	2.772711789109471	2.3215114515056423
Participant 32	1.7772480254849385	1.420311848332793

Table B.8: k-NNs results' barplots with VAE for the DEAP dataset. SVR achieves slightly better results from k-NNs, although k-NNs achieves almost similar with RFs.

	RMSE	MAE
Participant 1	2.495497376593442	2.312626811594203
Participant 2	3.0371909843306377	2.6321702898550727
Participant 3	1.4901570020216757	1.2486920289855072
Participant 4	2.079441709434616	1.8098623188405798
Participant 5	2.4119145439848864	2.039594202898551
Participant 6	1.6780554949617705	1.475865942028986
Participant 7	2.2172224790092834	1.9360253623188413
Participant 8	1.85288453316922	1.5023840579710146
Participant 9	1.4160204343548142	1.1676702898550726
Participant 10	1.8301311759811227	1.524163043478261
Participant 11	2.4058322729662334	1.9809347826086956
Participant 12	2.1896859451613837	1.910032608695652
Participant 13	2.539467586994137	2.1253442028985505
Participant 14	2.1653515240596204	1.9271992753623186
Participant 15	2.198387748084099	1.8551847826086956
Participant 16	1.9614397850165441	1.6279782608695654
Participant 17	1.2547554093790947	1.012855072463768
Participant 18	1.2664632418815702	1.0405289855072464
Participant 19	1.8466193591685198	1.579605072463768
Participant 20	1.585552508759652	1.2715797101449275
Participant 21	1.8998193012604419	1.5830000000000002
Participant 22	2.0820471199074264	1.6822282608695651
Participant 23	2.1363306634340735	1.853061594202898
Participant 24	2.000449604473406	1.6626521739130435
Participant 25	2.204952826246751	1.7364927536231882
Participant 26	2.454783688079563	2.0416014492753622
Participant 27	2.1316037842002054	1.5819818840579714
Participant 28	2.664958090797843	2.396985507246377
Participant 29	2.630543958687733	2.317079710144928
Participant 30	1.4088673858685505	1.1179963768115941
Participant 31	2.630366147745125	2.2103333333333333
Participant 32	1.787389981015448	1.4931847826086955

Table B.9: RFs results' barplots with VAE for the DEAP dataset. SVR achieves slightly better results from RFs, although RFs achieves almost similar with k-NNs.

	RMSE	MAE
Participant 1	2.555167639168003	2.270469109195402
Participant 2	3.057720169678831	2.6457966954022987
Participant 3	1.5071151149271231	1.2509619252873563
Participant 4	2.177047495257901	1.907819683908046
Participant 5	2.41847595077864	2.0277571839080455
Participant 6	1.7761841594278833	1.526433908045977
Participant 7	2.190926941501683	1.8466989942528735
Participant 8	1.8552872382330925	1.5060474137931037
Participant 9	1.3877482305366167	1.156622126436781
Participant 10	1.7107717478512836	1.4083793103448274
Participant 11	2.360309980271696	1.970931752873563
Participant 12	2.254593629181045	1.9417356321839077
Participant 13	2.6659977508673567	2.245935344827586
Participant 14	2.205787822722617	1.9151551724137932
Participant 15	1.3720833359662616	1.084340517241379
Participant 16	1.9956702196274385	1.604019396551724
Participant 17	1.2386606875223807	0.9675510057471266
Participant 18	1.2153386647884639	0.9928462643678158
Participant 19	1.9473164034291934	1.6242241379310343
Participant 20	1.5517654307045723	1.2319446839080461
Participant 21	1.9370590366882565	1.5886975574712647
Participant 22	1.9931446990328374	1.689007902298851
Participant 23	2.056513651301276	1.7127140804597698
Participant 24	1.86433034040722	1.4840043103448277
Participant 25	2.1448608757973	1.7357816091954021
Participant 26	2.2650433993270016	1.8800811781609192
Participant 27	2.141902851643121	1.5594439655172418
Participant 28	2.575636919414499	2.2226206896551726
Participant 29	2.6787371055088713	2.349244252873563
Participant 30	1.4095014489604052	1.1182744252873564
Participant 31	2.610751985766384	2.1752341954022993
Participant 32	1.7852440226124502	1.477199712643678

Table B.10: SVR results' barplots with VAE for our dataset. Overall results for our dataset outperformed those of DEAP. It is shown that SVR achieves the best good results from all three algorithms.

	RMSE	MAE
Participant 1	1.8868585036694043	1.1747280846661592
Participant 2	2.099584455562767	1.4757895928048754
Participant 3	2.0789657135880977	1.3797546848213254
Participant 4	1.800932494633672	1.3373896527552824
Participant 5	1.3130910568334961	0.9051263619293884
Participant 6	1.8352606459766931	1.0871705118294819
Participant 7	2.3719531500269864	1.4335345263323316
Participant 8	1.6636867415954002	1.039549778200753
Participant 9	1.9010528425028563	1.4127763922912677
Participant 10	1.7527511598390684	1.1317913104875441

Table B.11: k-NNs results' barplots with VAE for our dataset. Overall results for our dataset outperformed those of DEAP. SVR achieves slightly better results from k-NNs, although k-NNs achieves almost similar with RFs.

	RMSE	MAE
Participant 1	1.777727736754	1.3148037159124961
Participant 2	1.9615275107524226	1.6457543859649126
Participant 3	1.9614530954407337	1.5564912280701755
Participant 4	1.6971455912486522	1.4405614035087726
Participant 5	1.2780753073383702	0.9882105263157893
Participant 6	1.7238602133258134	1.22680701754386
Participant 7	2.221278910897954	1.7127017543859646
Participant 8	1.5915072851220287	1.1365614035087717
Participant 9	1.8176831214005813	1.4630175438596495
Participant 10	1.7100076946579725	1.3372631578947367

Table B.12: RFs results' barplots with VAE for our dataset. Overall results for our dataset outperformed those of DEAP. SVR achieves slightly better results from RFs, although RFs achieves almost similar with k-NNs.

	RMSE	MAE
Participant 1	1.7885167541711222	1.34127486662509
Participant 2	1.9742720327341432	1.6544646722607743
Participant 3	1.9988520927978422	1.5706912389588306
Participant 4	1.7290763309265154	1.4458898039527412
Participant 5	1.2836857231015257	0.9923560241354679
Participant 6	1.7422591296926884	1.2251323596930666
Participant 7	2.2248291106689084	1.7006857334818921
Participant 8	1.5998009136517382	1.1676963524252826
Participant 9	1.8440467963858258	1.4640634338510117
Participant 10	1.7095253166524598	1.3344914061826938

Appendix C

Additional plots

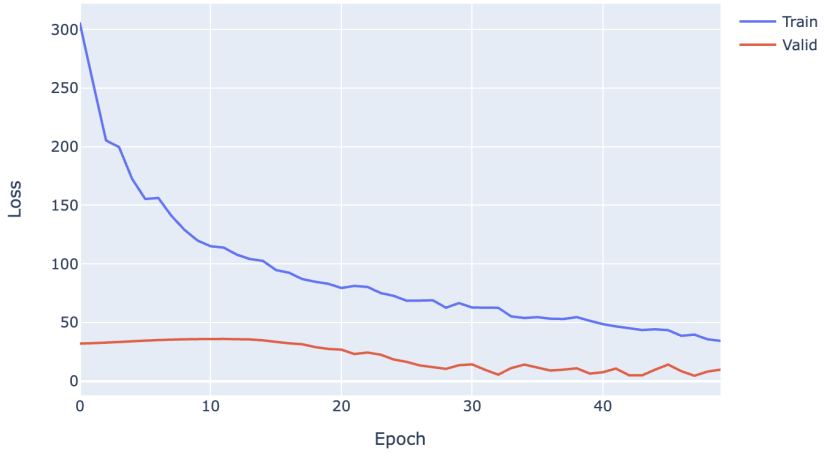

Figure C.1: Loss function visualisation for VAE for the DEAP dataset. It clearly indicates that throughout epochs the validation loss is smaller than the training loss, meaning that it's not overfitting and both losses are small meaning that it's not underfitting.

Appendix D

Hyperparameter tuning plots

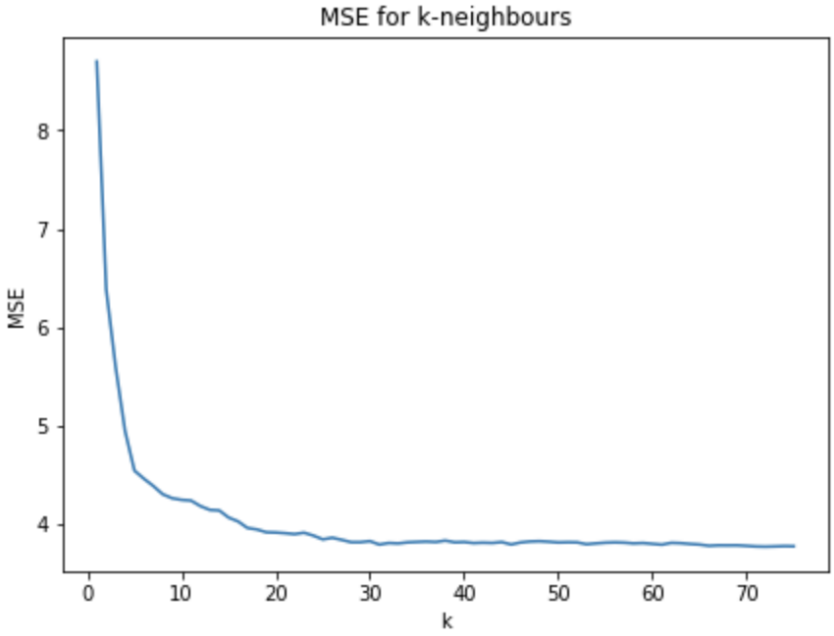

Figure D.1: Tuning for k neighbours in k-NNs with VAE on our dataset .Computing the MSE for separate models while increasing k we observe that the optimal k is 71.

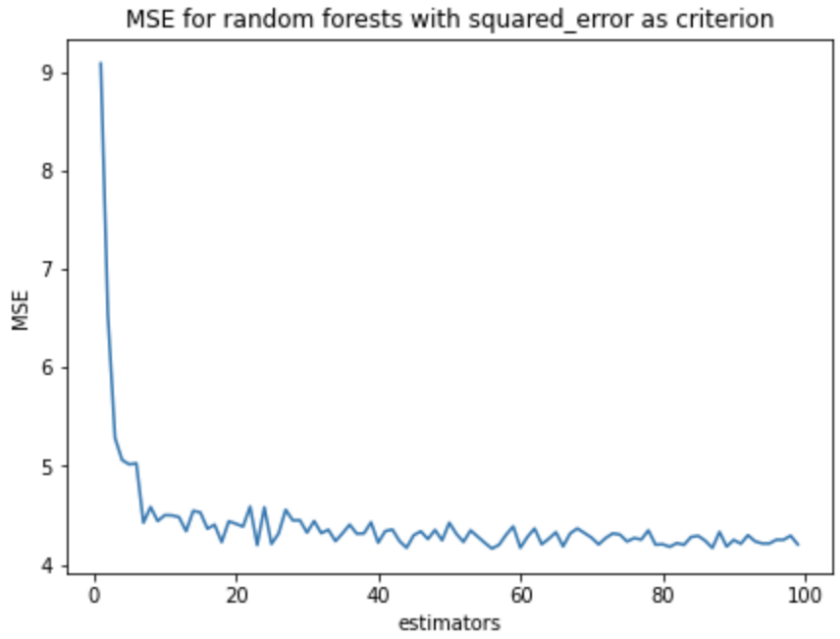

Figure D.2: Optimal number of decision trees for RFs using VAE on our dataset. Computing the MSE for separate models while increasing decision trees we observe that the optimal number is 55.

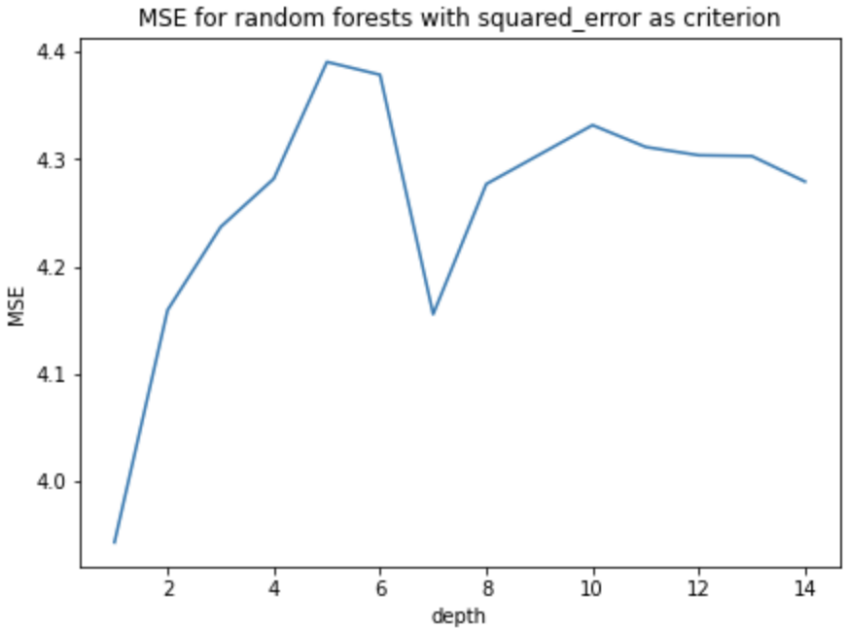

Figure D.3: Optimal maximum depth of decision trees for RFs using VAE on our dataset. Computing the MSE for separate models while increasing the depth we observe that the optimal maximum depth is 1

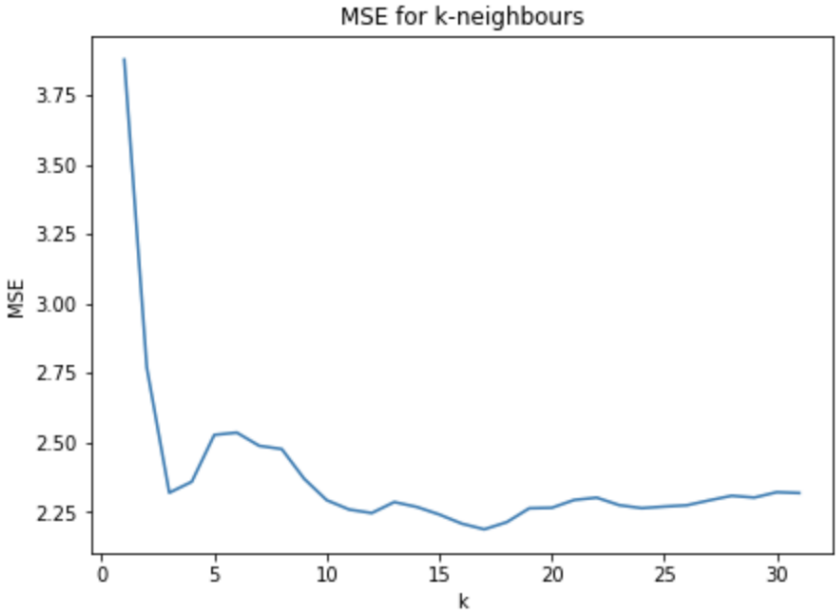

Figure D.4: Tuning for k neighbours in k-NNs with ICA on the DEAP dataset. Computing the MSE for separate models while increasing k we observe that the optimal k is 16.

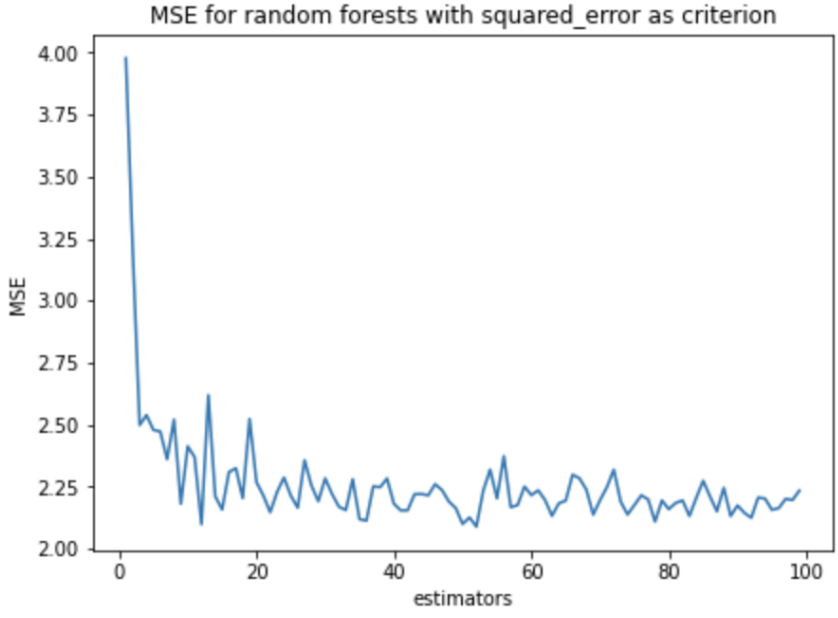

Figure D.5: Optimal number of decision trees for RFs using ICA on the DEAP dataset. Computing the MSE for separate models while increasing decision trees we observe that the optimal number is 51.

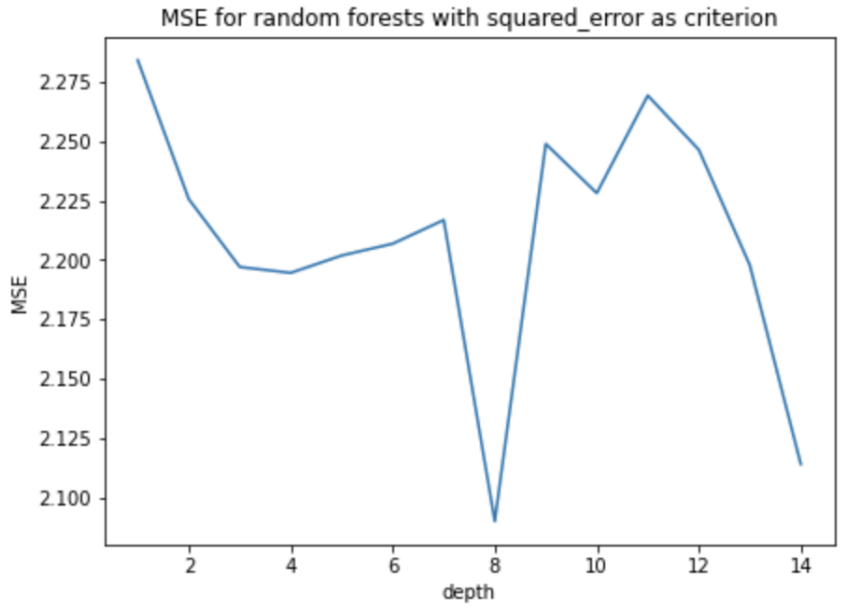

Figure D.6: Optimal maximum depth of decision tress for RFs using ICA on our dataset. Computing the MSE for separate models while increasing the depth we observe that the optimal maximum depth is 1.

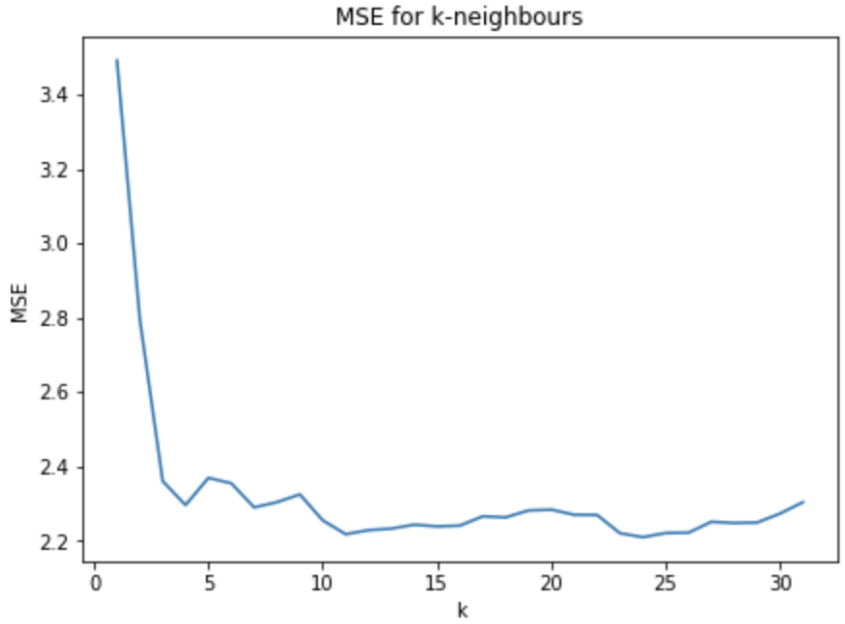

Figure D.7: Tuning for k neighbours in k-NNs with VAE on the DEAP dataset. Computing the MSE for separate models while increasing k we observe that the optimal k is 23.

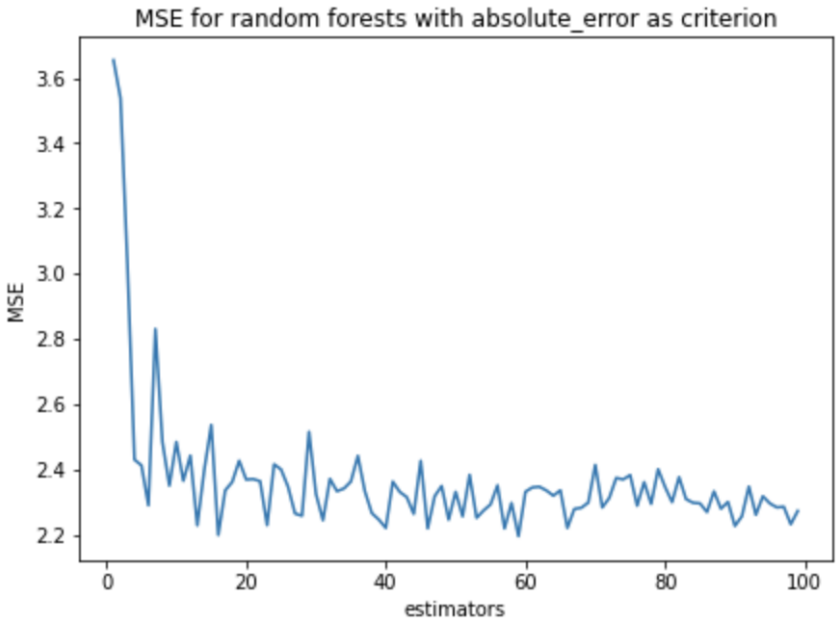

Figure D.8: Optimal number of decision trees for RFs using VAE on the DEAP dataset. Computing the MSE for separate models while increasing decision trees we observe that the optimal number is 58.

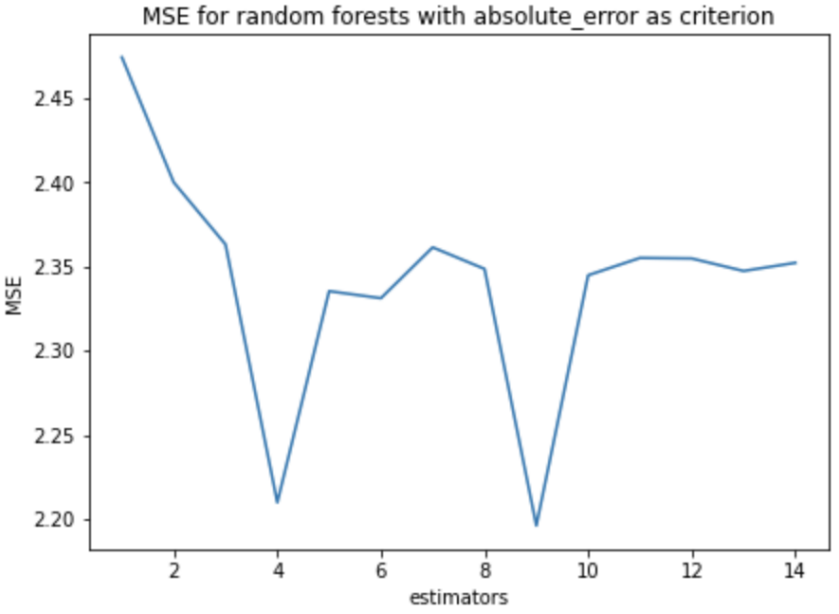

Figure D.9: Optimal maximum depth of decision tress for RFs using VAE on the DEAP dataset. Computing the MSE for separate models while increasing the depth we observe that the optimal maximum depth is 9.