BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF INFORMATION TECHNOLOGY
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF COMPUTER SYSTEMS
ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ

EMOTION RECOGNITION FROM BRAIN EEG SIGNALS

ROZPOZNÁVÁNÍ EMOCÍ Z EEG SIGNÁLŮ MOZKU.

MASTER'S THESIS
DIPLOMOVÁ PRÁCE

AUTHOR **Bc. KAREL FRITZ**
AUTOR PRÁCE

SUPERVISOR **Doc. AAMIR SAEED MALIK, Ph.D.**
VEDOUCÍ PRÁCE

BRNO 2022/2023

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Master's Thesis Assignment

141164

Institut:	Department of Computer Systems (UPSY)
Student:	**Fritz Karel, Bc.**
Programme:	Information Technology and Artificial Intelligence
Specialization:	Machine Learning
Title:	**Emotion Recognition from Brain Electroencephalogram (EEG) Signals**
Category:	Biocomputing
Academic year:	2022/23

Assignment:

1. Study and learn about the mood and emotion and how they affect the brain in terms of brain signals and images.
2. Get acquainted with signal & image processing methods as well as machine learning techniques and their application to the brain EEG signals and images.
3. Find out the challenges in recognition of mood & emotion from brain signals and images as well as the limitations of the existing methods.
4. Design an algorithm for recognition of mood & emotion from brain EEG signals and images.
5. Implement the designed algorithm.
6. Create a set of benchmark tasks to evaluate the quality of recognition of mood & emotion as well as the corresponding computational performance and memory usage.
7. Conduct critical analysis and discuss the achieved results and their contribution.

Literature:

- According to supervisor's advice.

Requirements for the semestral defence:

- Items 1 to 4 of the assignment.

Detailed formal requirements can be found at https://www.fit.vut.cz/study/theses/

Supervisor:	**Malik Aamir Saeed, doc., Ph.D.**
Head of Department:	Sekanina Lukáš, prof. Ing., Ph.D.
Beginning of work:	1.11.2022
Submission deadline:	31.7.2023
Approval date:	31.10.2022

Abstract

This study targets classifying emotion states, from Electroencephalogram (EEG) signal. Combining knowledge about physiology of the brain (and emotions), with frequency analysis, complexity analysis, signal processing and deep machine learning (CNN, GNN). Goal of this work is to create the emotion classification model and provide new insights into emotion recognition from EEG. Models created stands on the principles of CNN, GNN, multitask and self supervised training. One of the results achieved State of the Art results on the SEED dataset. Sharing process of understanding this task at the end of the thesis.

Abstrakt

Tato studie se zaměřuje na klasifikaci emocí z elektroencefalogramu (EEG). Kombinuje znalosti o fyziologii mozku (a emocí), s frekvenční analýzou, analýzou složitosti, zpracováním signálů a hlubokým strojovým učením (CNN, GNN). Cílem této práce je vytvořit model pro klasifikaci emocí a poskytnout nové náhledy do rozpoznávání emocí z EEG. Vytvořené modely stojí na principech CNN, GNN, multitask a self supervised tréninku. Jedním z výsledků bylo dosažení State of the Art výsledků na datasetu SEED. Proces porozumění této úloze sdílím na konci této práce.

Keywords

EEG (Electroencephalogram), Deep learning, Neural networks, emotion recognition, emotion classification, machine learning, brain, multitask learning, self supervised learning, graph neural networks, contrastive learning, convolutional neural networks, gradient propagation

Klíčová slova

EEG (Electroencephalogram), Deep learning, neuronové sítě, rozpoznávání emocí, klasifikace emocí, strojové učení, mozek, multitask učení, self supervised učení, grafové neuronové sítě, kontrastivní učení, konvoluční neuronové sítě, propagace gradientu

Reference

FRITZ, Karel. *Emotion Recognition from Brain EEG Signals.* Brno, 2022/2023. Master's thesis. Brno University of Technology, Faculty of Information Technology. Supervisor Doc. Aamir Saeed Malik, Ph.D.

Emotion Recognition from Brain EEG Signals

Declaration

I declare that I have prepared this diploma thesis independently, under the supervision of Doc. Aamir Saeed Malik, Ph.D. Most of the additional information was provided by Dr.Soyiba Jawed, MSc, and Ing.Michal Hradiš, Ph.D. I have cited all the literary sources, publications, and other resources I used.

.....
Karel Fritz
July 31, 2023

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Doc. Aamir Saeed Malik, Ph.D., for his unwavering support and guidance throughout the course of this thesis. His expertise and mentorship have been invaluable to me, and I am grateful for the opportunity to learn and grow under his tutelage.

I would also like to extend my sincere thanks to Dr. Soyiba Jawed and Michal Hradiš, Ph.D. for their substantial contributions to my research. Their insights and suggestions have played a critical role in the development of this work.

Special thanks to my girlfriend, who has been my constant source of motivation and encouragement. Her unwavering belief in me and her patience during the stressful times have made this journey easier. Her love and support have been my strength in times of hardship, and for that, I am eternally grateful.

Finally, I wish to express my heartfelt gratitude to my family. Their endless love, support, and encouragement have sustained me throughout my academic journey. Their faith in my capabilities continues to inspire me to reach for the stars.

I dedicate this work to all of them. Without their unceasing support and belief in me, this thesis would not have been possible.

Contents

Chapter 1

Introduction

Emotion recognition, some people do have ability to recognize how the person feels, what emotional state the person currently experiencing. The characteristics of the recognition here are that it is surely based on previous experiences and probably on seeing the emitted facial expression of the subject. So we can say that sometimes emotions can be externally recognized. This means that we can adjust our behavior, as a feedback to mentioned emotion recognition. And so it is useful for us to be able to recognize different emotional states. It can be also useful for others, imagine person experiencing hard times, surely it can be helpful if others treat that person differently, for example calming the person down in case of anger.

Figure 1.1: Facial expression of different emotions [9]

Now, the question is, are we able to distinguish different emotions based on another source of information. One of possible sources could be from dynamic electromagnetic field surrounding, and also being generated in, the brain. Currently there are many solutions targeting solely problem of measuring this electromagnetic field. In case of this work, the measuring technique is called Electroencephalogram (EEG). An non-invasive technique able to capture potential changes on the surface of the head. Experiment procedure here is the crucial thing that matters. After that, we usually get the data hopefully containing the information sufficient for proper emotion recognition. The way we get the information from the data is the objective of this thesis. Experimentation nature of this thesis is scheduled as follows. First, we try to get maximal possible information from well known standard feature extraction methods. Then the complexity will slightly increase when trying to find optimal deep learning neural network models for enhancing the information gain even more. At the end, the final ensemble is going to be build. Which could have the best performance on emotion recognition from EEG, and which is the goal of this thesis.

Chapter 2

Understanding the data

2.1 Brain

The brain is the most complex functioning system that we know. It's mostly made up from neuron cells, axons, glial cells and blood vessels. Brain is also our fattest organ consisting of approximately 60% fat. The surface of the brain is crumbled, that allows brain to increase the surface area, which is desired. Supreme layers of the brain consist mainly of gray matter (neuron bodies), therefore increasing the surface area could led to higher intellect capacity. Power generation of brain could light up a bulb and generation of thoughts per day is at about 70000 count. But above all, the brain is the seat of intelligence, interpreter of the senses, initiator of body movements, and controller of behaviour. Although only thanks to brain we can feel pain, the brain as such has no pain receptors. Anatomy of this incredible structure is one of the key knowledge domain needed for future work in this field. Because of that it's necessary to mention gross and micro anatomy.

2.1.1 Gross anatomy of the brain

Covering (Meninges) of the brain is made up from 4 main layers. First is skull, it's the hard cover, which should protect brain from outside dangers e.g. punch. Underneath the skull there is dura mater, that has two sublayers. The periosteal layer (cranium) and the meningeal layer (lower). Spaces between the layers allow for the passage of veins and arteries that supply blood flow to the brain. The underlaying arachnoid is just connective tissue and does not contain blood vessels or nerves. Below arachnoid, there is a fluidum called cerebrospinal fluid, which cushions the brain and also contain blood vessels. Last layer above the brain is called pia mater, it is rich in with veins and arteries.

Moving on to brain structure itself, from the most broad point of view brain can be divided into 3 parts, cerebrum, cerebellum, brain stem. Cerebrum is responsible for all higher order functions like thinking, planning movement, hearing, speaking and more. Closer to the surface, there is grey matter and more inside is white matter. Grey matter consisting of neuron bodies (soma), and white matter represents axons. Cerebellum is fist size part of the brain, which is involved mainly in coordination, complex movement would not be possible without this structure. Recent studies found out other roles which Cerebellum could have, these are for example emotions, social behavior or addiction. Despite of the size cerebellum contains more than 50% of all neurons in CNS (central nervous system). Brain stem is the oldest part of our brain responsible mainly for primary function like breathing and hearth

beat control. Due to that, damage to this part could be much more severe than damage to upper parts of the brain. If person loses a part of cerebrum, there is still possibility to survive, even a possibility of normal life! On the other hand, damage to brain stem could led to loss of heart beat control, which is fatal.

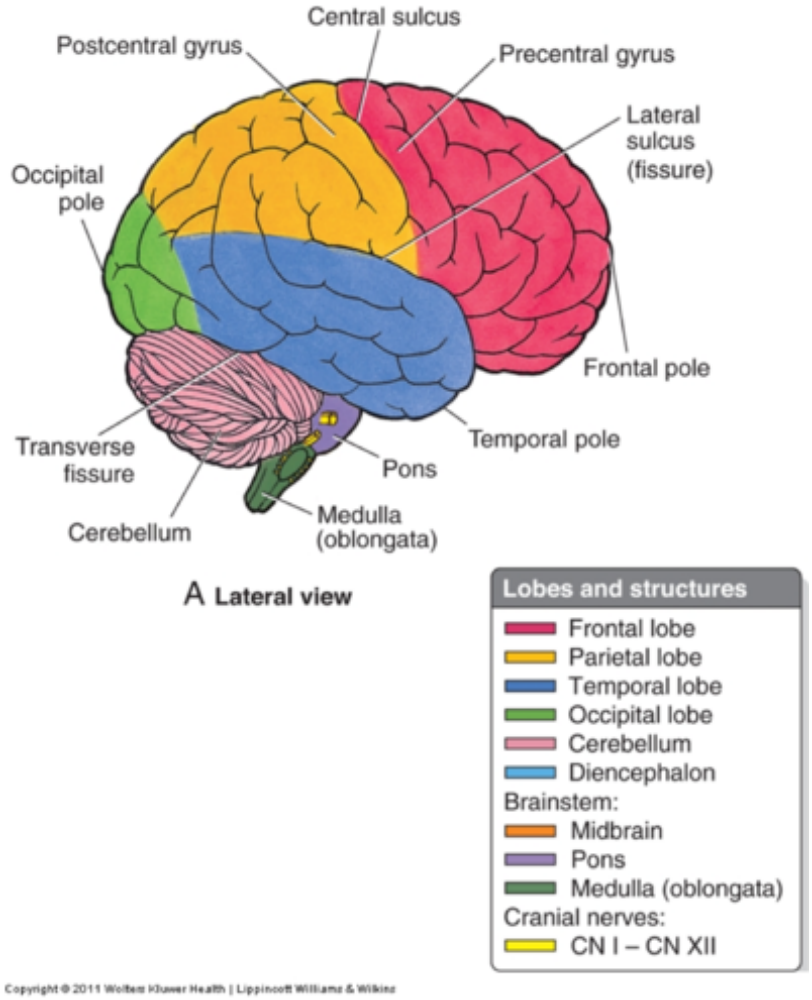

Figure 2.1: Brain Gross organization [13]

Brain stem is consisting of inner part of the brain called midbrain, next there is the medulla and pons. Midbrain, the old part of the brain serves perfectly as a connection between forebrain and hindbrain. Pons, links brain to the spinal cord, and manage to control breathing, sleep-wake cycles. Pons can also be considered as merging point for several cranial nerves, or as a bridge between two parts of CNS (central nervous system), spinal cord, medulla and upper brain. Medulla is at very bottom of brain stem, responsible for primary functions like heart beat. Medulla is also responsible for many reflexive reactions.

The covering of the midbrain, and the newer part of the brain is called Diencephalone. Diencephalone consists of Thalamus, Hypothalamus, Epithalamus, Subthalamus. Thalamus is the hub like structure, in which nerve fibers project to cerebral cortex in all directions. Hypothalamus is responsible for regulating certain metabolic processes and also controls hunger, body temperature and more. Moving to the next layer, which is cerebral cortex,

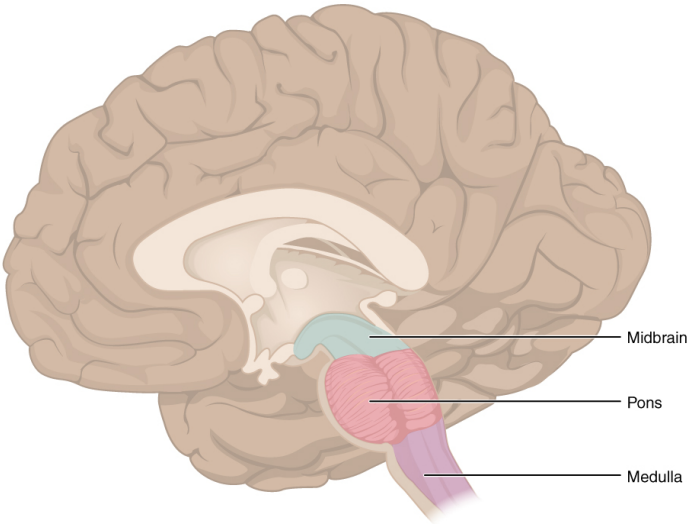

Figure 2.2: Brain Stem [3]

newest and most interesting part of our brain. Cerebral cortex has more than one way of decomposition, but the most used is into lobes. There is Frontal, Parietal, Temporal and Occipital (in the literature, there is sometimes a fifth lobe called central, this lobe is in between frontal and parietal).

FRONTAL

The frontal lobe is the largest lobe, and is associated with many mind like functions. This lobe contains most of the dopamine neurons. Dopamine controls reward, attention, motivation or planning. Information that is flowing from thalamus is therefore filtered and selected. Dopamines are the mechanism which allow our cerebral cortex to limit these informations.

There are many functions associated with frontal lobe, for example personality, problem solving, emotional expressions, attention, self-monitoring or motor control. Frontal lobe also involves ability to predict the future consequences of some decision or of current action. So we can talk about behavior control and planning. Talking about recognizing and classification of emotions, the frontal lobe could be good source of features, it is because this lobe is more related to emotion processing than other for example parietal lobe. But this is a question which this thesis is about, in later chapters I will do analysis of information contained in each lobe so I can then distinguish which is better for emotion classification than the others.

PARIETAL

The parietal lobe process information from various modalities. Modalities I am talking about are spatial sense, somatosensory sense (touch, temperature, and pain). Thanks to parietal lobe we can sense size, shape or color. Reading, writing and math skill have also lot to do with the parietal lobe. For emotion classification, parietal lobe should be less informative, but this needs further analysis.

OCCIPITAL

Occipital lobe is responsible for processing most of the visual information (reaction to visuals, and also interpolation). For this thesis, I predict lower importance, because the lobe is not primary related to the emotions.

TEMPORAL

This lobe is important, it process emotions and feeling in some way. It is also a language center, learning and memory associated structure. Long term memory is produced via communication between temporal lobe and hippocampus. Temporal lobe is also commonly associated with processing auditory information. Damage to this region could led to aggressive behaviour or understanding the spoken word.

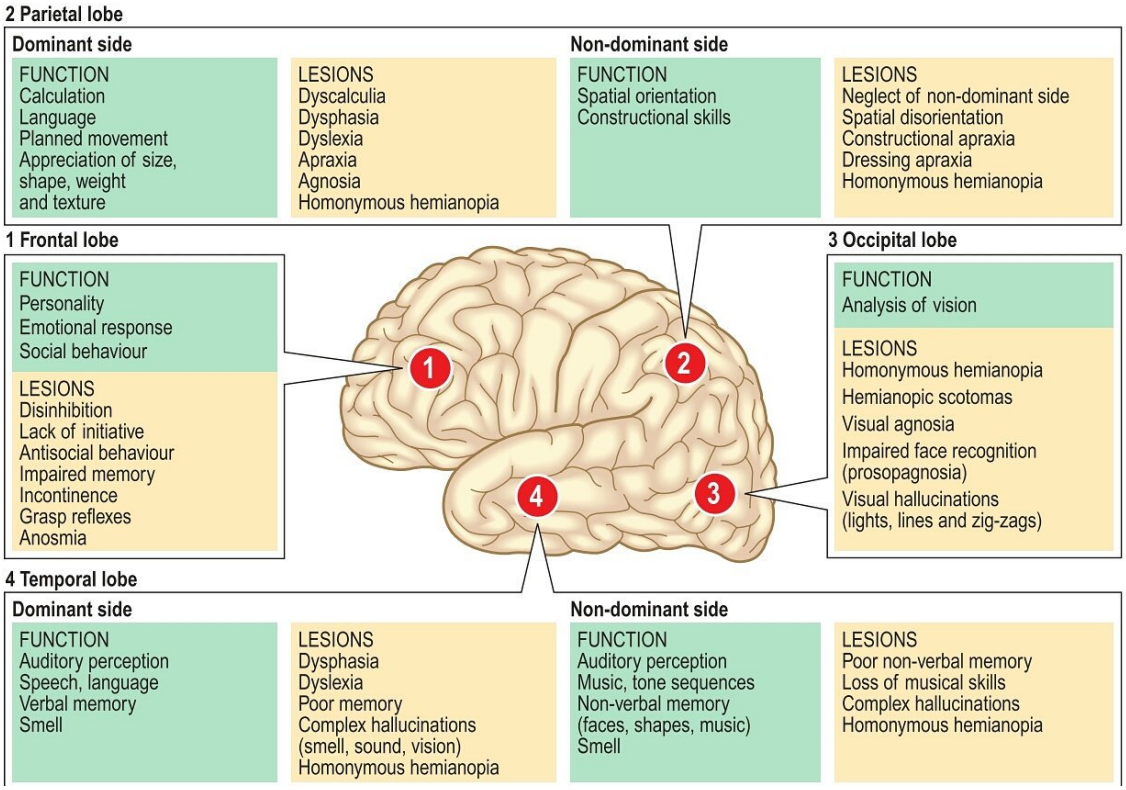

Figure 2.3: Brain lobes function [2]

Another thing that should be mentioned is that brain are 2 functioning objects. Those two objects are called hemispheres. Each hemisphere controls the opposite side of the body, that means if you want to move left hand, the signal will originate from right hemisphere. Hemispheres are connected with structure called Corpus callosum. Corpus callosum therefore allow the communication between those two hemispheres, and it is a largest white matter structure in a brain, there is approximately 200-300 millions of axonal projections [24].

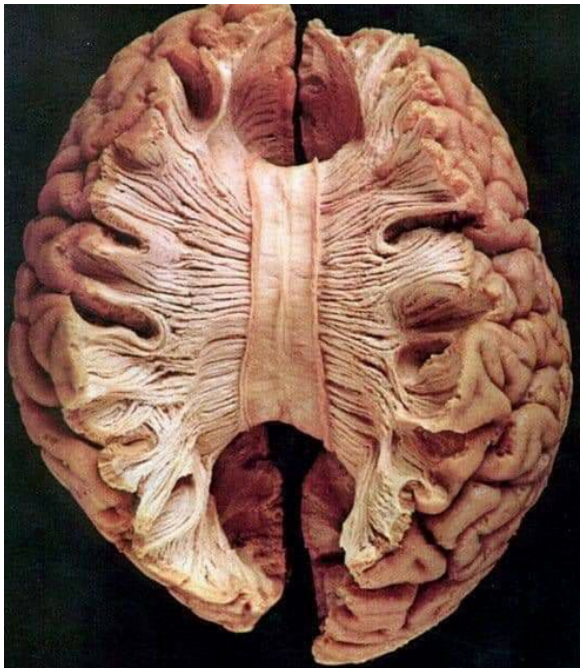

Figure 2.4: Corpus Callosum [6]

2.1.2 Microanatomy of the brain

Previous section talks about gross anatomy of the brain, this section is about anatomy at the lower level, the cellular level. The cell that is responsible for all of this is called neuron. Neuron is just like normal cell, despite of that it has mechanisms that allow neuron to process information. There are more than 200 types of neurons, these are deployed throughout the brain in very specific way. One of the most popular type, neocortex pyramidal neuron, is also located near the surface of the brain. In electroencephalography (EEG), we measure voltage changes primary thanks to clusters of pyramidal neurons.

A neuron is also called nerve cell. It consists of cell body (soma) and nerve fibers for communication. Fibers are called dendrites, one fiber is elongated forming a new structure, called axon, that serves for sending information outside of a neuron. Axon is covered in myelin sheath, lipid-rich substance. Myelin sheath allows faster rate of electrical potential

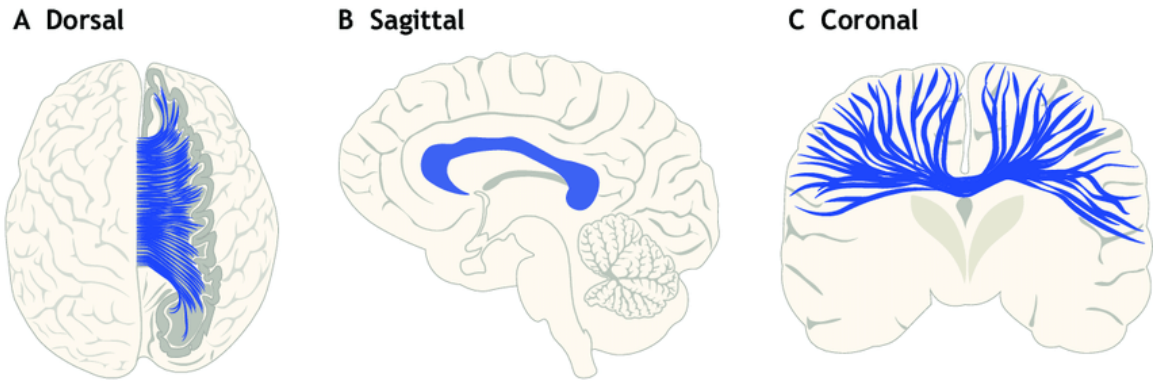

Figure 2.5: Corpus Callosum 2 [7]

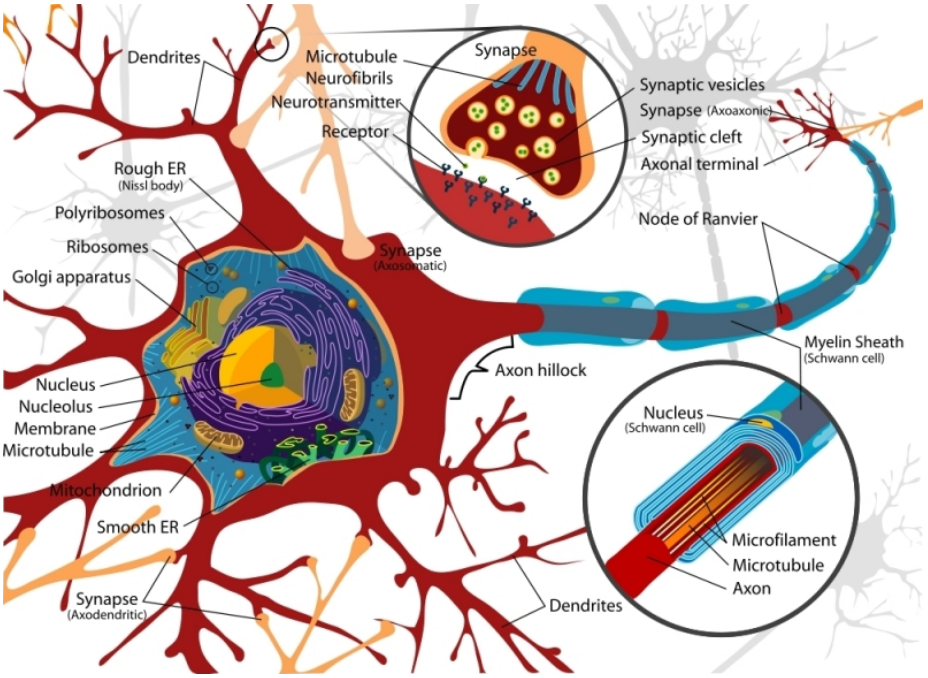

Figure 2.6: Neuron approximate model [17]

being transported along axon. When the end of axon meets another neuron, the place is called synapse and it is the place where the complex chemical-electrical reactions happen. Axon conducts action potential to the end, and there a specific amount of things called neurotransmitters are released into synaptic cleft.

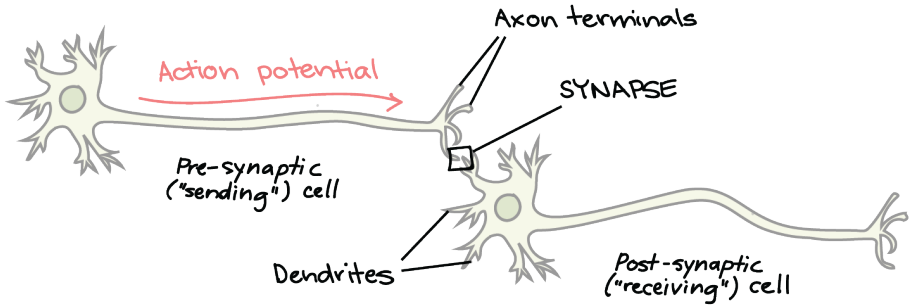

Figure 2.7: Synapse [1]

Neurotransmitters are signaling molecules, which somehow affects the receiving neuron. There are more than one hundred of different neurotransmitter types. For example dopamine, our NT of reward, or serotonin NT of happiness (and many more than just happiness). These transports of neurotransmitters in between end of axon of one neuron (also called pre-synaptic) and dendrite of other neuron (post-synaptic). Transportation of NTs into another neuron can cause two possible reactions, one is inhibitory and the other is excitatory. Excitatory simply is trying to get the receiving neuron into action, and inhibitory is doing the opposite. If receiving neuron receive enough excitatory NTs through his dendrites, it does an event also called emitting action potential. Process which I described, can now be repeated, since the action potential simply means that neuron fire another electrical

potential along axon, targeting next neuron. It is important to mention that brain diseases we know, cannot be well grasped without the knowledge of these principles. For example parkinson disease is caused by lack of dopamine production.

2.2 Emotions

Feelings, thoughts or behavioral responses, or degree of pleasure/unpleasure [36] these all are part of term emotions. Mental states which are result of brain activity. Emotions are also associated with terms like mood, temperament, personality, disposition or creativity [26]. Emotions can result in various physiological, behavioral and cognitive changes. Higher order emotions like love, hate, happiness we can experience were much more basic back in the past. The original function of emotion was to recognize danger situations and behaviour leading to survival and reproduction [60]. Emotions are fairly complex thing to get good grasp on, it is still not clear if cognition is an important aspect of emotion or not. This is surely important question in terms of doing analysis of more than one participant, since their neural expression to the same emotions could be various based on cognition capacitance. Another questions targeting whether or not emotions cause changes in our behaviour [53].

As another example of how complex this topic is, consider extrovert people versus introvert people. These two groups have diametrically different way of living an emotion. Extrovert people tend more to expresing their emotion and being more social than the introverts, introverts tend to live the emotions inside. Does this play a role in EEG as well? Recent studies have shown that it does, not so much in alpha frequencies when eyes closed, but it starts to be significant when considering theta and beta waves when eyes closed. On the other hand differences in alpha waves can be detected more easily with eyes closed. Across all frequencies together, extroverts yield higher output, making them more active in term of neural-voltage power. These studies didn't went deeper into a specific patterns, which if they are the same in an extrovert and introvert, could allow us to classify emotions independently on the personality.

Different point of view, understand emotion as a reaction to environmental stimuli [53]. Most important characteristic in this case is our awareness. Generaly understood as reactions to memories, ideas, or actual happening in near environment. This point of view also suggest that bad emotions lead not only to bad psychological state, but also to bad physical state, people make decision based on emotions. Commonly positive emotions are considered as healthy, on the other hand negative emotions could cause worse heath and living standards [56].

2.2.1 Representation of emotions

There are many ways how to represent emotions, the main categorization is into two types: Discrete [35] and Continuous or also called dimensional type [62]. Discrete one categorizes emotions to multiple main or most well known emotions. The most simple discrete categorization is into positive and negative emotion. There are many discrete models for representing emotions. Most proposed theory based on Darwin [33] and Tomkins [68] says that discrete categorization comprises of nine basic emotions. Those are interest-excitement, surprise-startle, enjoyment-joy, distress-anguish, dissmell, fear-terror, anger-rage, contempt-disgust, and shame-humiliation. And it is believed that these nine play

represent an image of mental health of the individual. Mr Ekman has came out with another well accepted theory, and that is, each essential emotion must follow 4 basic rules. First, emotions are instinctive, that means you are not having them because you want to. Second, various people generate the same emotion in the same situation. Third, various people express these emotions in comparable way. And the fourth, the last one, physiological patterns of different people are similar when experiencing basic emotions. According to Ekman, sadness, surprise, happiness, disgust, fear and anger are examples of basic emotions, these are even detectable solely from facial expression. Other theorists came out with different discrete models, which some of the emotions considered in these models can be seen in the picture 4.2, where each row represents different model.

Emotions
Surprise, joy, interest, rage, disgust, fear, anguish, shame
Fear, sadness, happiness, anger, disgust, surprise
Rage and terror, anxiety, joy
Pain, pleasure
Fear, love, rage
Fear, grief, love, rage
Expectancy, rage, fear, panic
Sadness, happiness
Anger, courage, aversion, dejection, despair, desire, fear, hope, hate, sadness, love
Happiness, sadness, fear, anger, disgust
Desire, interest, happiness, surprise, sorrow, wonder
Anger, disgust, contempt, distress, guilt, fear, interest, shame, joy, surprise
Anger, fear, elation, disgust

Figure 2.8: Discrete emotions used in different discrete models [43]

The problem with a discrete models is that it is believed they have several limitations in terms of representing specific emotion. Means they are not able to describe the wide range of emotional states. In other words, emotional states are too complex for representing them with few discrete emotions. Dimensional models try to tackle this issue by categorizing emotions continuously in the space, usually 2D space. Each axis in this case represents an emotional characteristic. Emotion is therefore a point in this multidimensional space.

As some examples of multidimensional space, there is Russell's arousal and valence 2D model [63]. It can describe up to 150 distinct emotions. Second example could be Whissell's model [70], which again is two dimensional. One dimension represents evaluation scale and other the activation. And the third example could be Schloberg's model which is unusually 3D, adding attention-rejection dimension.

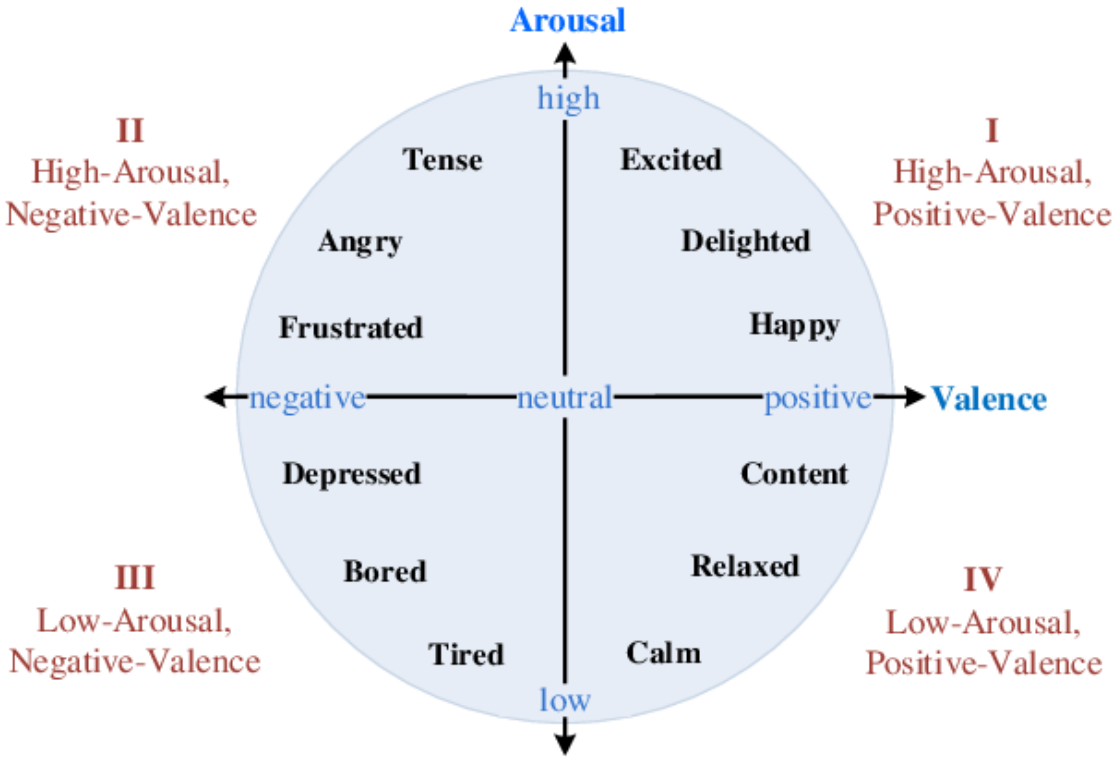

Figure 2.9: Two Dimensional valence-arousal emotion model [22]

Most used continuous model for representing emotions is the one from Russell, so the valence-arousal one. The valence dimension shows joy, reactiveness, cheerfulness or sadness of emotion, ranging from negative to positive ones. Arousal on the other hand represents the intensity of the emotion, ranging from low to excitement. So there are 4 quadrants, negative low intensity emotion, negative high intensity emotion, positive low intensity emotion, and positive high intensity emotion. First, negative low intensity quadrant localised on the left bottom, contains emotions like sad, bored, or sleepy. Second, the negative high intensity emotion represents nervous, angry or annoying emotions. Third, the positive low intensity quadrant contains emotions like calm, peaceful or relaxed. And the last fourth quadrant contains the positive high intensity emotions such as pleasure, happiness or excitement.

Russell's model is used often in analysis, and it is also used in datasets like DEAP. DEAP also uses models like-dislike and dominanace-familiarity. SEED dataset uses positive/negative/neutral discrete model and SEED-IV happy/sad/neutral/fear. DREAMER dataset is using the three dimensional model of valence, arousal and dominance. This model can be seen in picture 2.10, where there are the six basic emotions plotted, according to Ekman.

2.2.2 Emotional recognition

After all, why do we want to be able to classify emotions solely on EEG? So many answers to this question, as many fields associated with human-robot interaction are booming. And still, the most advanced systems nowadays cannot understand the emotional part well.

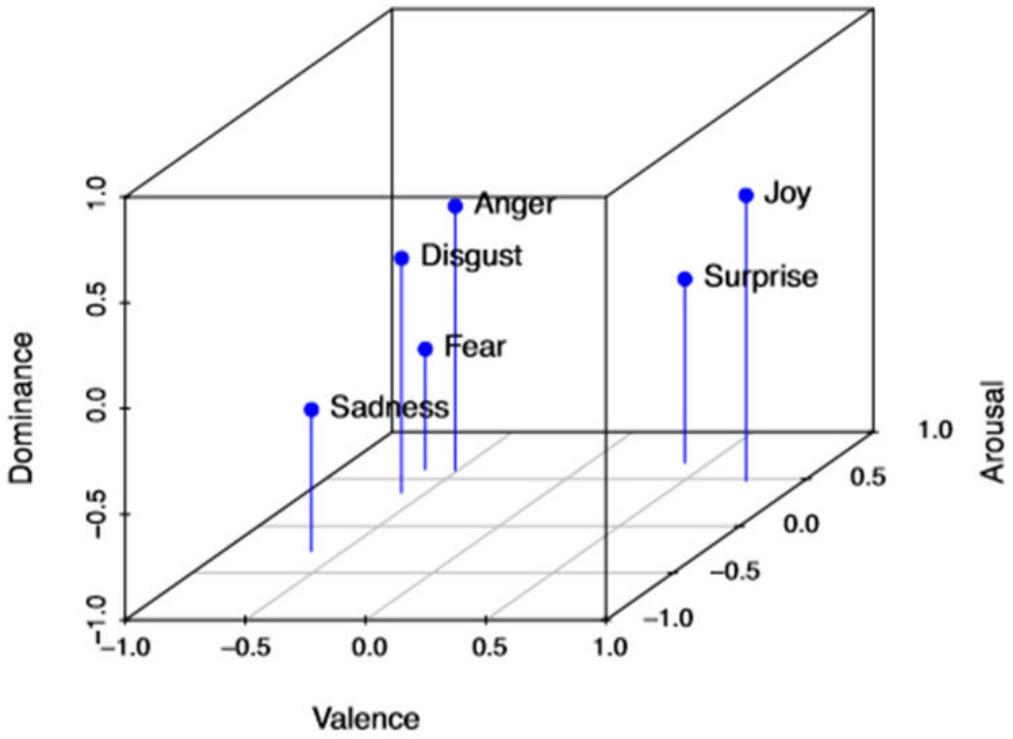

Figure 2.10: Three Dimensional valence-arousal-dominance emotion model [21]

They are incapable to distinguish between various emotions and then take a decision based on that. Emotions can be recognized from various sources, for example facial expression, or even EKG signal, to distinguish if person is nervous or not. Those advances, hopefully will lead to better and more user friendly systems.

First step is to gather an emotional response through EEG, then to recognize it. For eliciting and emotional response, there are several ways to do that. One way is to elicit emotion based on simulated scenarios. If person is recalling what happened in the past, usually it comprises of the emotion associated with this memory fragment. This approach is easy and can be varied a lot, that is useful when doing analysis of a single person. But if we want to average, each person can elicit different emotion based on just recalling experiences and memories.

Second approach uses photos, videos or sounds to evoke desired emotions, as this is less error prone in terms of gathering desired emotion, it is therefore also more frequently used. Third option is to play a game, which carry an big advantage. Person can experience the emotional more intensively, that is because playing is more like experiencing, then just watching a listening to passive stimuli. Strongest elicitation was measured in participants, who played games such as flying simulator [69], and that is logical, since the simulation offers most of the authenticity and rawness from real life.

There are several sources for emotion elicitation. For sound, there is Affective Digitized Sound System (IADS) [71], or its expanded version IADS-E [72]. This dataset contain

Figure 2.11: Examples of emotional visual stimuli [10]

many sounds for each topic, there is for example 56 breaking sound or 64 electronic sounds and so on. For pictures, there is International Affective Picture System (IAPS) [31]. Same as previous this dataset contain images that are able to evoke similar emotions, in different people. IAPS is made up of twenty groups, each of size 60, so there is 1200 pictures in total. Each picture in IAPS has its valence and arousal values, in the IADS, there is valence, arousal and dominance score. Both of those datasets are label by human, which can cause inconsistencies and need a future study. This hopefully will ensure better generalization over BCI (Brain computer interface) applications.

2.2.3 Brain recording modalities

When trying to get and information from data, we need the context of the data and our knowledge. Therefore we should know what exactly is an Encephalogram measuring (EEG). Measuring technique called EEG is an non-invasive method, which is able to measure the brain activity with a very good temporal resolution and not so good spatial resolution. It measures the activity of whole brain, but the most precise measuring is done on the neocortex zone. This neocortex measuring is due to his location, it is right underneat skull and meninges, it is a first most top layer of brain tissue as such.

Number of other possible methods is possible. Their main differences are in form of experiments (and their cost, or portability), temporal precision and spatial precision. Most well known methods are fMRI, MEG, fNIRS, PET, MEG, ECoG, electrode arrays and finally EEG.

PET Positron emission tomography (PET) scan, has for example much better spatial resolution than EEG, but much worse temporal resolution. Its principle is based on injecting a radioactive substance called tracer into a human body, this substance can then be observed. What is beautiful about this method, is that it shows us a bit about brain and its tissue functioning. This is an advantage, other scanning methods, often gives information only about structure of the brain. Main use of Positron emission tomography is to detect

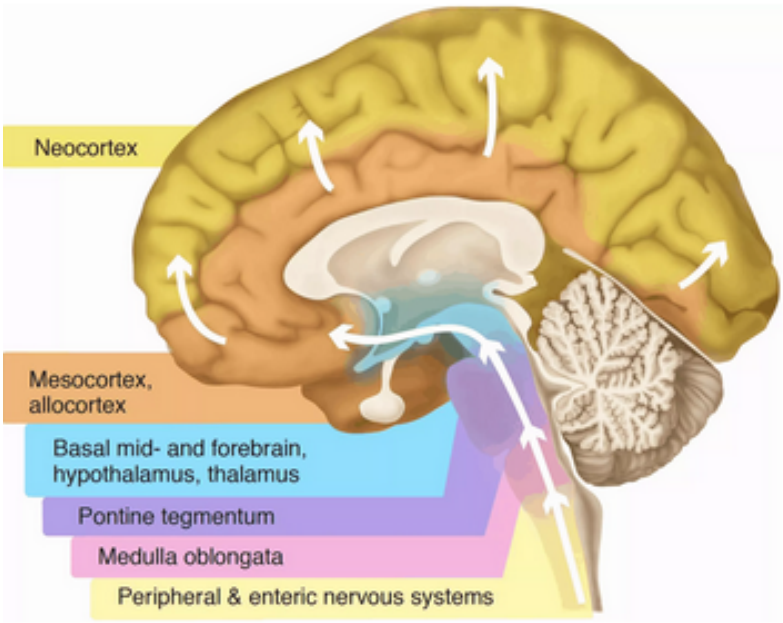

Figure 2.12: Brain internal organization

diseases inside the brain, via measuring brain metabolism and distribution of radioactive substance throughout the brain. Main disadvantage of PET is bad temporal resolution and also its portability or cost.

fNIRS Functional near-infrared spectroscopy (fNIRS), is another technique for measuring brain activity non-invasively. Alongside EEG, it has similar characteristics, its has better spatial resolution and slightly worse temporal resolution, and again it is fairly portable. It is based optical brain monitoring technique. Neuroimaging is done using near-infrared spectroscopy. Its results are often compared with fMRI, but fNIRS is only able to measure regions near the cortical surface [23].

ECoG When doing EEG experiment (or measuring), channels are placed right on the surface of the head. On the other hand, ECoG channel are measuring the brain activity with channel as well, but now the channels are below the skull. Thanks to being closer to the brain, ECoG can obtain much better spatial resolution, also not loosing any of the temporal resolution. These are the advantages, which are indeed very good. The disadvantages about ECoG are mainly its invasiveness and the limited point of view. Limited field of view means that the method is limited by the measured area of the cortex.

Electrode array (implant) The most invasive, at this time, is the straightforward implants. These implants directly penetrate the brain tissue, which allows this method to get the best overall signal-to-noise ration. Brain implants are currently being used, just in clinical application, for example to treat parkinson's disease patients. Or another application currently used is to determine what happen to the brain after some type of injuries. For example how are the specific areas damaged, after undergoing a stroke. This is an overall common purpose nowadays, because of the riskiness. An exception are animals,

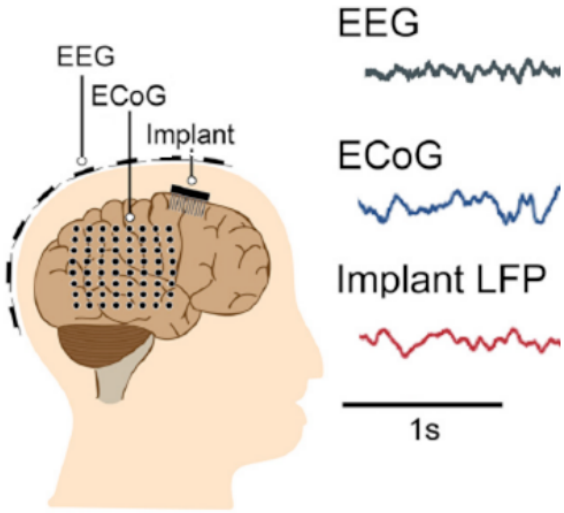

Figure 2.13: EEG vs. ECoG vs. Implant [8]

implants are commonly used to measure their brain activity for scientific reasons. Some brain implants are even used for BCI purpose, it is possible, but the big disadvantage is the invasiveness, risk, cost, limited field of view (accessible cortex area is limited), and the no less important environment deterioration over a time. Environmental deterioration means that over a time brain creates a matter around the implant causing the impossibility of further measuring, there will be too high impedance.

According to [19], these are the specific temporal and spatial precision of these methods.
EEG: spatial resolution (7-10 mm), temporal resolution (<1 ms)
MEG: spatial resolution (2-6 mm), temporal resolution (< 1 ms)
fMRI: spatial resolution (1-2 mm), temporal resolution (1s)

2.2.4 EEG, experiment

The data of an electroencephalogram experiment are gathered through special devices which can measure the changes of potential on the outside of a skull. Therefore measuring via EEG is an non-invasive method, which is big advantage. The changes are usually very small, ranging from -70 to +100. Because they are so small device for measuring must be placed properly so the data are not meaningless (good Signal to Noise Ratio). One of the most important thing is the design of the experiment. Designing the circumstances, for example person is just sitting and trying to lift and object, could bring us a lot more of information about what is happening in the brain when this specific movement is being planned or being executed, surely more than just measuring the date without these rules. Another big advantage is that EEG is highly portable, for example against MEG, PET and others where you need a specialised room for experiment.

Montages

When measuring the brain signals, montages are specifying how to place the electrodes correctly. Some well known montages are 10-20, 10-10 or 10-5 systems. For example the

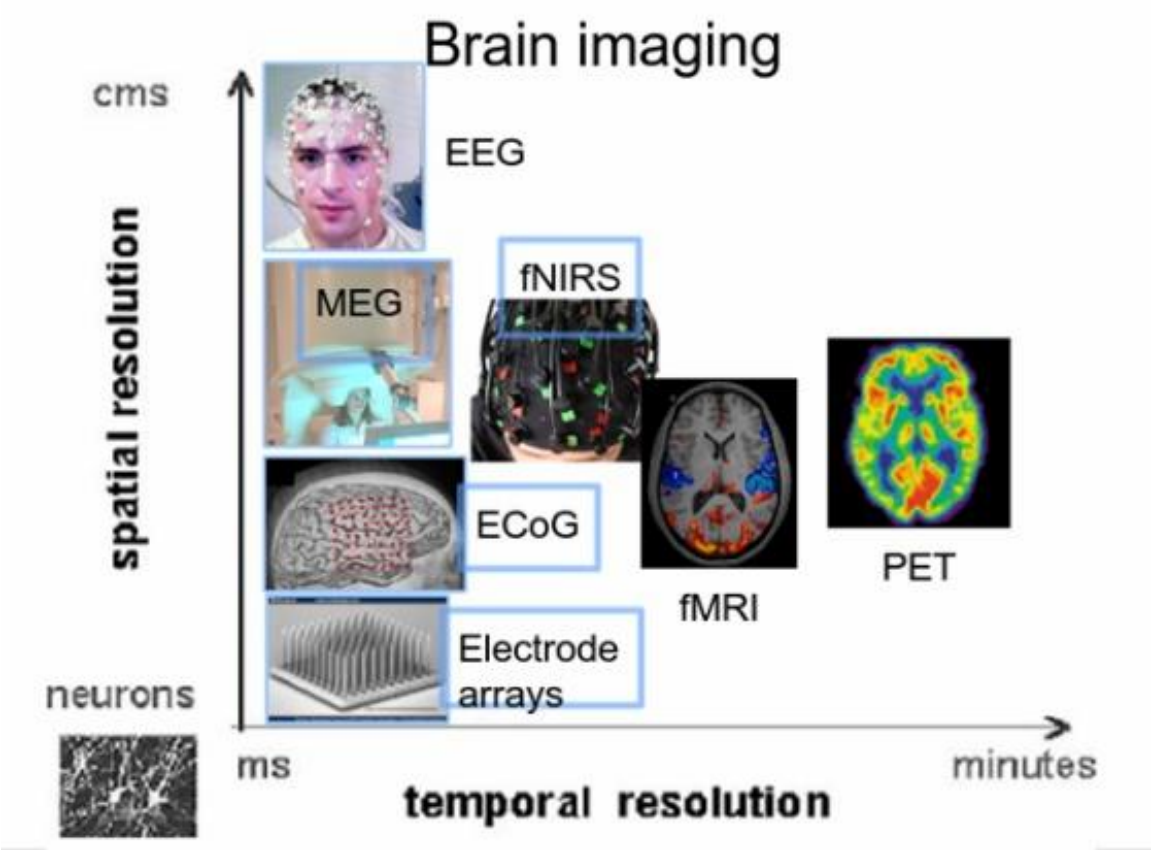

Figure 2.14: Temporal / Spatial performance [19]

10-20 system arrange electrodes relatively into increments of 10% or 20%, ensuring that each electrode is relatively positioned to all the others. This makes experiments consistent for every EEG study despite many different head shapes and sizes. In frontal part the montage is based (start) on region called nasion (point between the forehead and the skull) and on the back of the head there is inion (bump at the back of the skull). Surface between inion and nasion is the area where electrodes will be placed.

Cleaning, Preprocessing and Filtering

Data from electroencephalogram (EEG) are very sensitive to any noise, it can be either line noise from radio, or noise from beating heart (we are usually not interested in this signal information). Therefore the data needs a special care at cleaning and preprocessing. The noise in the recording can be classified into two classes „non physiological“ and „physiological“.

Non physiological noise

Line noise (50Hz or 60Hz), this type of noise originate from the radio signal. It can be easily removed from the data with bandpass (notch) filtering of the specific frequencies. This incorporates with an minimal information loss, since we are interested in activity of our „thinking“ we are interested only in frequencies in 1Hz-40Hz.

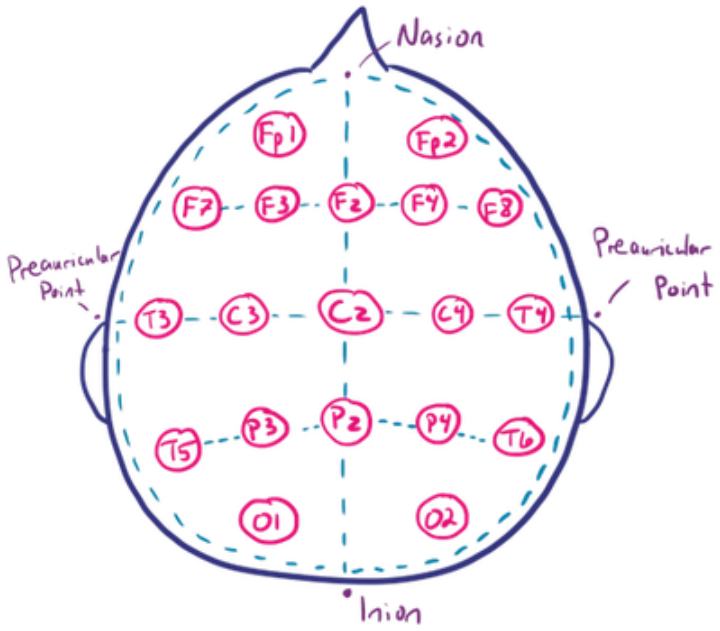

Figure 2.15: Montage 10-20 [16]

Bad setting of the electrodes, this can also cause problems, for example if one electrode unstick from the skin. Poor grounding of the EEG is similar problem, when the channel (electrode) have a high impedance, it can't measure well.

Physiological noise

Eye movement artefact, ECG artefact, EMG artefact etc. these all are noise which came from your body. It is not always easy to remove them, but number of useful methods are available, for example Independent Component analysis is good for removing eye moving artefact, Canonical Correlation Analysis is especially good for muscle artefact because of the form of the artefact. ECG can be removed same as eye and muscle noise, or can be removed based on recording aside cerebral, as a pattern [45].

As the brain activity is between 1Hz-40Hz, we usually remove higher or lower frequencies so they are not considered as effective in classification. Removing is done by just apply a bandpass filter.

Further Preprocessing

EEG signal has its standard formats, like edf, fif and more. Standard format guarantee easier manipulation, meta infromation about the recording, even the epochs for ERP (Event related potentials). So it brings us many advantages, unfortunately not all recordings are in this format.

Many datasets are in a form of collection of csv files, these can be loaded into pandas dataframe and or into mne format.

MNE

Open-source Python package for exploring, visualizing, and analyzing human neurophysiological data: MEG, EEG, sEEG, ECoG, NIRS, and more. We can load edf, fif etc. to mne object which automatically gives us a set of collection of methods such as PSD of the signals, plotting topomaps, build-in ICA, convert to numpy and many more.

The goal is to build a classifier, so the format must adapt these requirements. Numpy array is suitable as input so in the final data are going to be converted to numpy array.

2.3 EEG analyses for emotion classification

To tackle the problem of classifying emotions, people came with a typical pipeline for processing EEG data. The gathering of EEG data, was described in the previous chapter. The next step would be the cleaning and preprocessing the data, which is important because bad data means bad results. Cleaning usually means filtering and denoising (from artifacts). Preprocessing addresses normalization, segmentation and the very important feature extraction. All these processes are done to have a better and suitable input for a classifier.

2.3.1 Cleaning

Filtering

EEG data as such do not need very high and very low frequencies, therefore is a good practice to remove them. Brain activity is approximately from 1Hz-40Hz, but different literature says different values. Sure the filtering depends on the application, for example emotions, tends to leave most information in alpha and theta frequencies, on the other hand there is motor movement classification which omits more of lower frequencies. Most

used type of filtering in EEG is just band-pass to fixed frequencies. This means that the filter is specified via two values, which first is the minimum frequency and the second is maximum frequency. If we use this filter on the data, they will be deprived of <1Hz or >40Hz frequencies. The effect of this filtration can be seen in the picture 2.17.

In this example frequencies below 0.5Hz and frequencies above 34Hz are removed. Interesting peak can be seen on unfiltered data at 50Hz, this is not a natural activity nor bad channel placement, it refers to frequencies used by radio broadcast. Successfully removing these frequencies so they will not influence the future classification, and we will also get better results. There is lot of different strategies for even better filtering like adaptive filters and more, but for this case classical band-pass filter should be enough good.

Denoising

EEG data are specific type of recording burdened by error corresponding to noise from our own body and also the non-physiological noise. Getting rid of unwanted noise is usually done by a few methods that deal with different types of noise.

Eyes can generate a significant potential, which we measure, and which is not wanted. This is called artifact, more precisely eye artifacts. Eye artifacts can be eye moving and eye blinking both have specific pattern and their source is at the front of a skull.

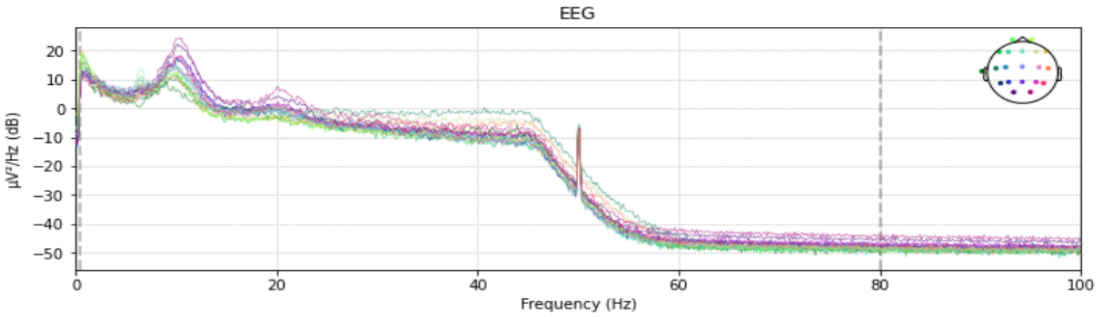

Figure 2.16: EEG data PSD before filtration

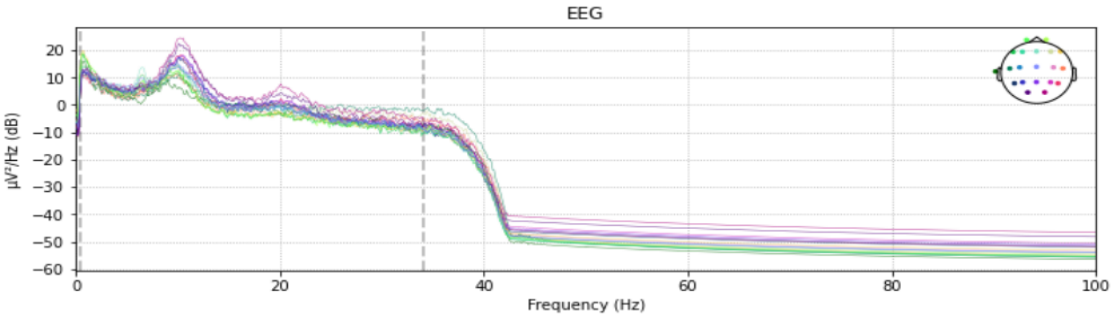

Figure 2.17: EEG data PSD after filtration

Another type of artifact is muscle artifact, it is due to muscle movement. We do not want these artifact unless we are interested in the movement, motor control or any analysis discussing the muscle movement. Maybe we want to measure just this (usually) noisy muscle signals, so we can further use it as a training examples of a classifier, which goal is to learn how to detect these artifact and remove them from original signal.

Denoising eye associated artifact can be done in number of ways, most popular is to use Independent component analysis (ICA). ICA will create n components based on the data, those components reflect the origin of potentials, there at least some of domain knowledge is needed. By visual checking, we determine which of components are eye artifact related and delete them. This process also loose some of the information, therefore is important to remove really the noisy artifact component. If the component is removed, it is substracted form each channel in original data. The example of running ICA on the data is shown in image 2.18. Clearly visible, some components have activated parietal lobe, some occipital and some frontal. Those in frontal, could refer to eye artifact, especially if the activity is sharp and in the front of the head. In the picture 2.18, the ICA component number 15 looks like containing eye artifact values, therefore it is going to be removed.

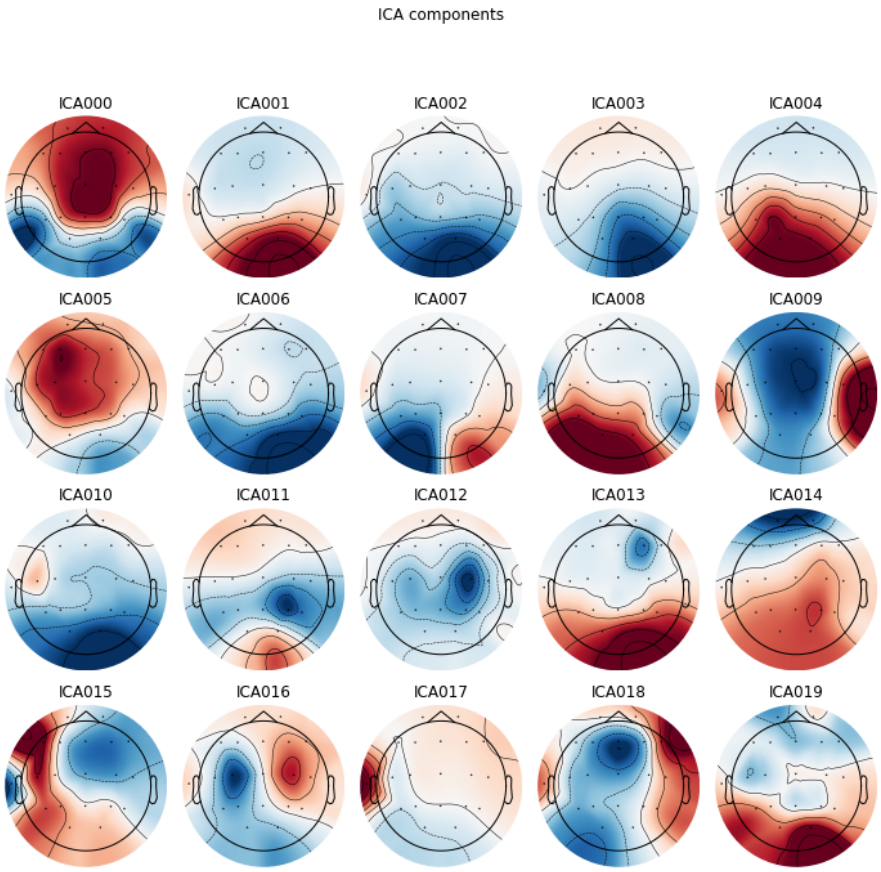

Figure 2.18: Independent component analysis results

Removing the muscle artifact will be done with method called Canonical correlation analysis (CCA). This algorithm aslso has the non trivial math behind. First it computes the correlation matrix and then a several math operations are applied, resulting in a muscle artifact removal method, which outperforms the ICA at these muscle artifacts removal. The complete description of this algorithm can be found in the original paper [[34]].
The application on noisy data can be seen in picture 2.19. Clearly it can be seen that those that looks like sequence of spikes, which are muscle artifact, are being successfully removed, hopefully not taking much cognitive information with them. Mark of success is that even after removal, the relatively higher frequencies, for example in channel Fp2 or F8 were kept.

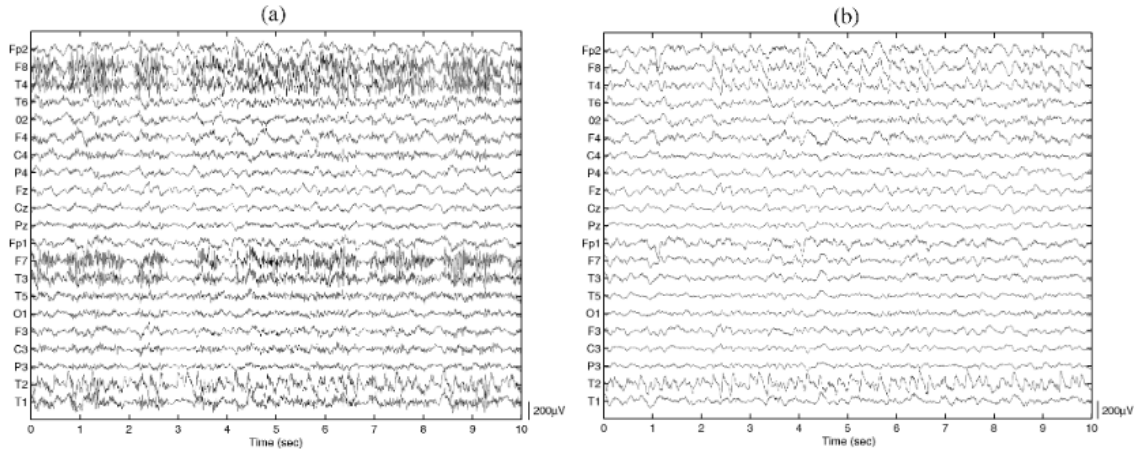

Figure 2.19: Canonical correlation analysis results [32]

2.3.2 Preprocessing

Preprocessing is a stage of analysis process, in which we are again, trying to get better quality inputs for our classifier. In other words, improving the quality of data, trying to get maximum from our data. This refers to terms normalization, segmentation, and the very most popular feature extraction, which will have an individual section.

Normalization

The normalization helps classifier to distinguish classes with less effort. Also tranforming the inputs to have common scale. This helps especially when two columns, which are the same importance. For example, education (if we encode education done into values 1-10) and salary (ranging from 0-100000 and more), then the salary will have more impact on the result, which is not something that we want. Therefore we normalize, in EEG most used type of normalization is the Z-transform, also called Z-score. The methods formula $z = (x - \mu)/\sigma$, tells us something about the value itself, and also where it lies in the distribution. This step is important because, after this all channels will contribute to the classification result equally, otherwise they might create an unwanted bias.

Segmentation

In many applications data segmentation of the signal processing task such as automatic analysis of EEG signal, is used to keep some information about time in EEG. There are more types of segmentation, the most general type is about segmenting signal with fixed length of chunk. For example the values are grouped by the 10ms interval, therefore if the sampling speed is for example 500Hz and chunk size is 10ms, in every group there will be 5 records. From these 5 records we can for example extract features which will classify the data better, not only because we are now classifying based on five time more data, but if we classifying only one record there is no information about transition over time which is no less important.

ERPs

The recording of EEG is often based on so called ERPs, which stands for Event Related Potentials. Experiments principle is that some stimulus is shown, or done to participant, and then the brain response is monitored. Very useful and important experiment design is we want to know the exact patters for some task. EEG is very noisy data, so ERPs strategy is to record many trials over many participants, and then average the recordings over all trials and all participants. This technique gives us way how to analyze the cognitive response to the stimulus, and also the sensory response. Sensory response appear around 100ms after the stimulus, if the time is shorter it could mean the superior mental performance on the other hand longer time may mean that the person has limited possibilities. Sensory response is hard-coded, it cannot be learned how to optimise this. The cognitive response doesn't share this characteristics, while learning or doing random things, the brain is trying to optimise the pathways effectiveness. This is the reason why cognitive response can vary if the person is taking steps to improve.

2.3.3 Feature Extraction: Source Localization

When measuring the EEG data, the only thing that we get is the changes of potential on the surface of the head. There is no information about what is happening deeper inside the brain. This question has no unique answer, it is an ill posed problem. Information about what is the source of the signal that is measured on the surface, is the result of methods called „Source localization". Goal here is to localise the individual signal sources.

There are few well known methods, for doing such analysis. These methods are are complex ones and will not be used in this work. Instead it is important to mention that this analysis exists, and could deliver promising results. Methods for doing such analysis are for example MNE (Minimum Norm Estimation), LORETA (Low Resolution Electromagnetic Tomographic Analysis), MUSIC or FOCUSS (Focal Under determined System Solution). All of them has own advantages and disadvantages, the use is depend on the task. Additionally there are hybrid methods and modification to previously mentioned methods.

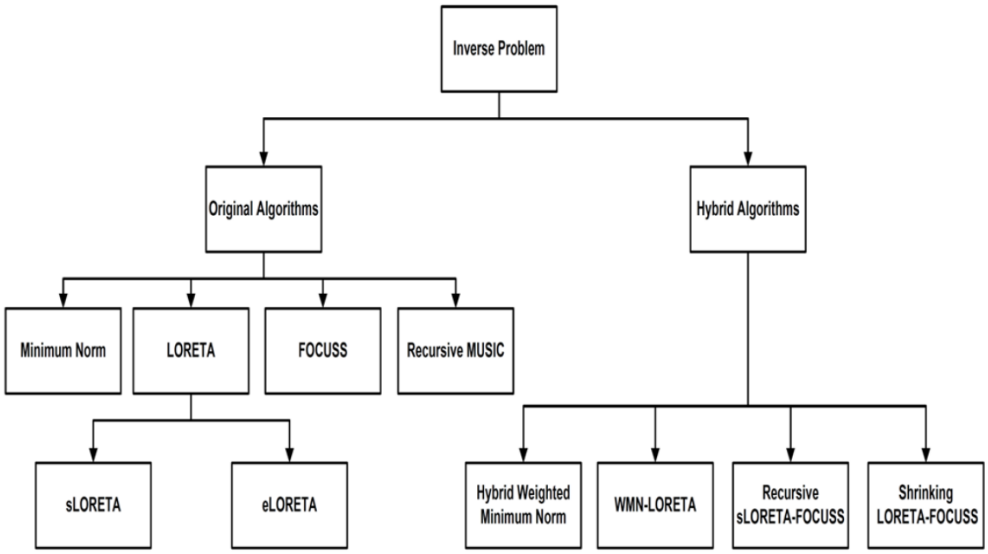

Figure 2.20: Inverse problem solutions

2.3.4 Feature Extraction: Frequency analysis

Frequency bands

Frequencies occurring in human brain can be categorized into five main categories. Sometimes there is one additional category called Mu, which corresponds to motor functioning. The lowest frequency band is the Delta band, it ranges from 0.5Hz - 4Hz and is associated with stages of deep dreamless sleep, or coma. These waves usually occurs in the frontal cortex. Theta waves can be found mainly in temporal and parietal lobe and they are associated with relaxed state. Theta band should contain information about emotion processing as recent studies shown, so it is interest of this thesis. Alpha waves are mainly associated with occipital and parietal. Detected in resting state with eyes closed. Alpha waves have higher oscillatory energy than beta or gamma in both positive and negative emotions. So they are too, more interesting for this thesis than higher frequencies or delta. Beta waves are typically observed in the frontal lobe, but they can occur in a variety of locations. Beta happens when the mind is active and focused. Gamma waves are found in variety of networks (sensory or non-sensory). Gamma can observed durning multi-modal tasks, that means that the task involves high level functions like reception, processing, integration, transmission, and feedback.

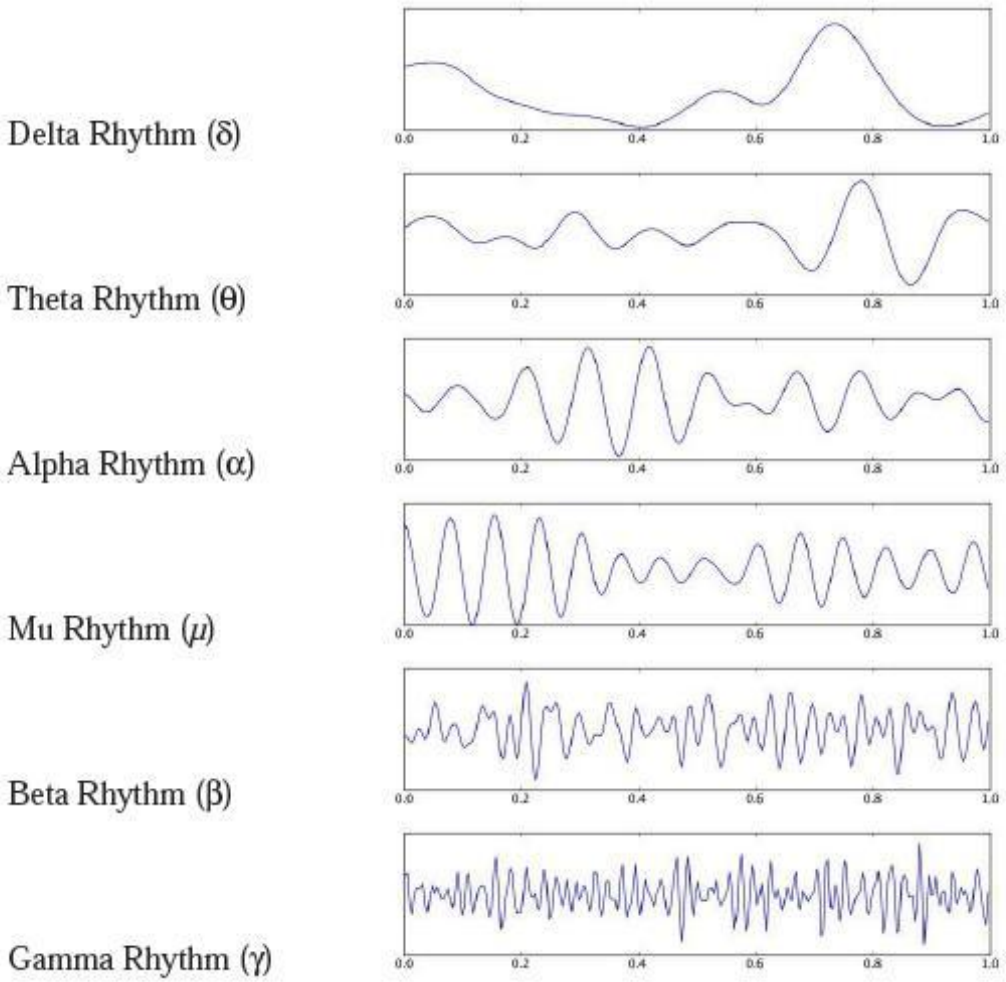

Figure 2.21: Different frequency bands [25]

Segmenting the signal based on dominating frequency band can be useful in further classification. The most discriminating features should come from frequency analysis of alpha and theta waves in the frontal and temporal lobe. Although these are the main areas of interest, the classification will be done on whole spectrum, every lobe, just filtered with bandpass filter 0.5Hz to 42Hz. If any of the lobe would contain more information, model should re-weight autonomously.

Peak Frequency, Peak Power, Bandwidth

Peak frequency and peak power both omits information about time, therefore is lost. One way to deal with this is to segment data for example to intervals of 10ms. Peak frequency is about the highest frequency in the given signal, similarly peak power is about the frequency which has got most power in given signal. Bandwidth talks about in which frequencies the given signal is, this can be further controlled by setting a threshold or thresholds to distinguish which frequencies will be tagged as present. These features will surely be used and tested for they information value in question of classifying emotions.

Total power, Spectral Edge Density

Total power could be used as feature and is calculated as the total spectral power in the power spectral density. It is measured in decibels. Spectral edge density is the frequency under which 90% of signal power is. Both could be used as features, if this is only one additional input for each, adding them could lead to better results without increasing computation complexity significantly.

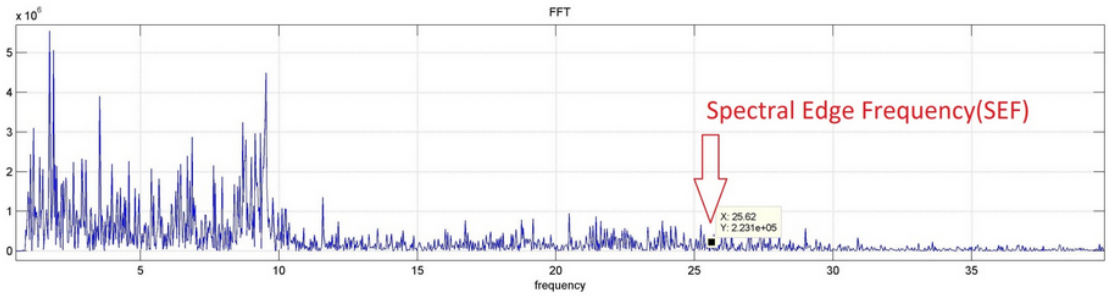

Figure 2.22: 90%

SEF method can be changed to any percentage resulting in differently valued results. Predicting that the SEF with higher values could be less informative, because emotions are happening more in lower frequencies.

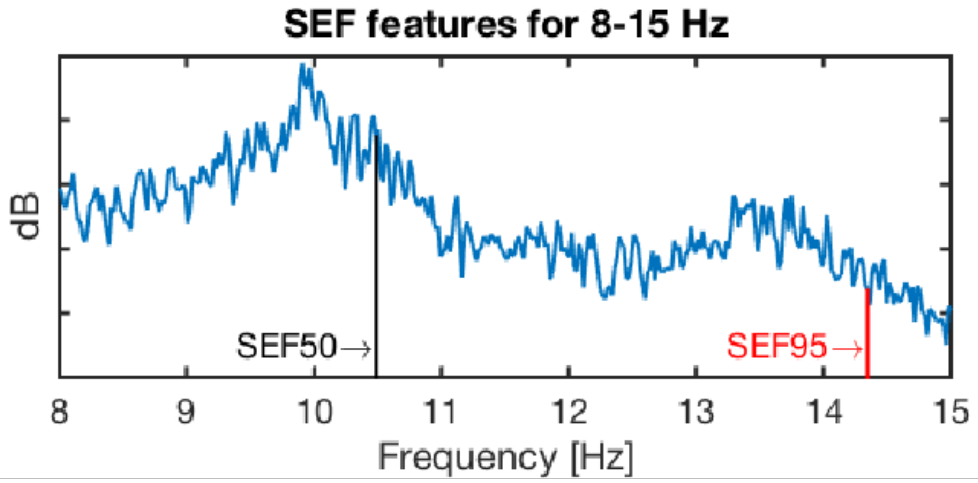

Figure 2.23: 50% and 95% [20]

Absolute and relative power

Value that represents power of certain frequency band independently of the other bands. This is called Absolute power and it is computed in Decibels. In the picture 2.24, we can see two channels and corresponding absolute powers of frequency bands.

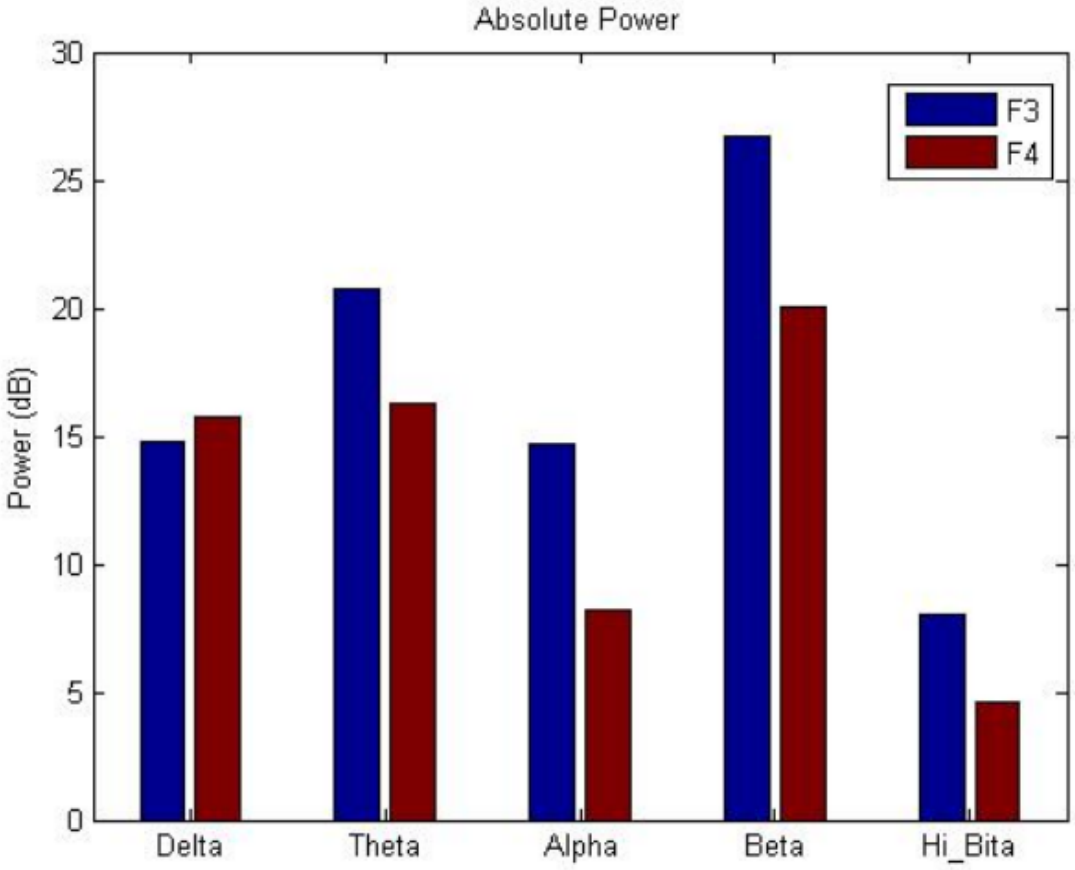

Figure 2.24: Absolute power (in Db)

On the other hand Relative power is using percentage instead of absolute value. The formula for calculating relative power is just the result from absolute power divided by total power and then multiplied by hundred, so it gives us a percentage.

Both of these techniques are valuable for classifier, amount of usefulness could be altered by adding different time information. For example calculating Absolute/Relative power in differently sized windows. Granularity of this solution is partly a trade of with accuracy. Future testing of this feature will be done and the appropriate proportion of it's information will be used. As relative inputs are normalized, they can be used as straight as inputs, absolute ones need to be normalized first.

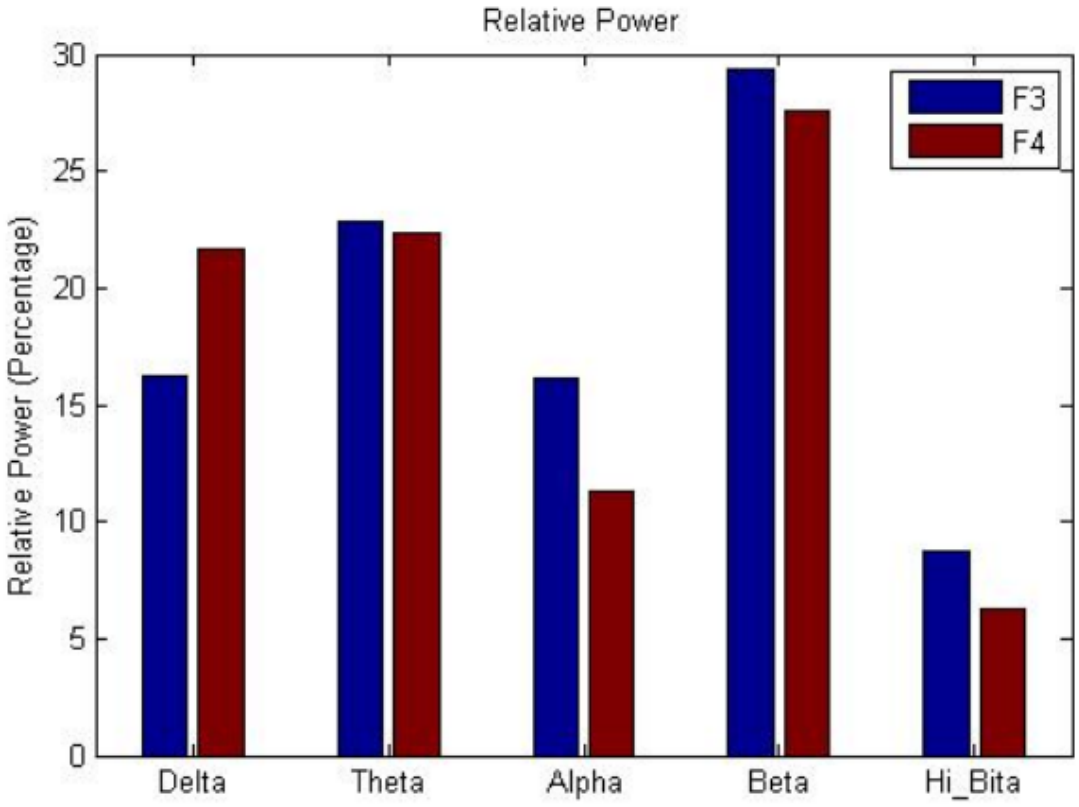

Figure 2.25: Relative power (in percentage)

Brain Symmetry Index

This method compares the absolute power of each hemisphere against the other. Value zero means perfect symmetry, value one then means maximum asymmetry. It is said that left prefrontal cortex could be the source of positive emotion processing, therefore asymmetry will again have an informative value for emotion recognition. Overall this method is somehow similar to absolute power of individual channels, only the precision is lower and less complex for further manipulation.

Amplitude asymmetry

Comparing the signal from two different electrodes. This method looks very simple, but it enables the great potential when choosing the right pair of electrodes. For example choosing electrode from one functional brain network and the other from different functional brain network, gives us an information about asymmetry between two electrodes in two different brain functional network. Can be varied by choosing not only two electrodes but maybe some groups of electrodes.

2.3.5 Feature Extraction: Time Analysis

Time domain analysis for EEG is comprised of techniques like Event Related Potentials, Entropies, Higuchi's Fractal Dimension, Hjorth Complexity, Hurst Exponent, higher-order

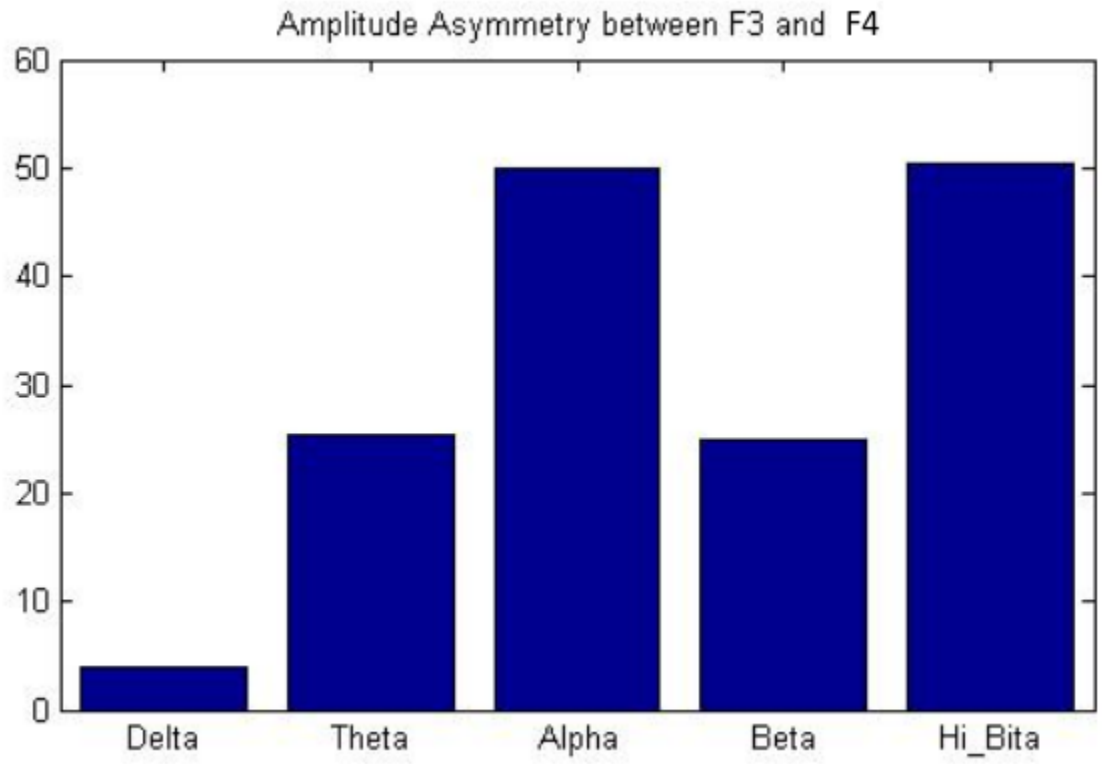

Figure 2.26: Amplitude asymmetry

crossing (HOC), Principal/Independent Component Analysis, self-similarity, activity, mobility, or Gray level Co-occurrence Matrix. In addition there are also the basic statistical features like mean, maxima/minima, median, standard deviation, skewness, relative band energy, kurtosis and more.

Entropies

Entropy is a measure of uncertainty, but it can express the complexity as well. Now, there is few different types of methods used for entropy feature extraction. Important thing while doing entropy analysis of EEG is the segmentation of original signal. For example if we have signal [4,2,0] and m=2, the first segment would be [4,2] and the seconds [2,0].

Sample Entropy Sample Entropy represents the signal complexity, where lower values indicates less complex signal (more self-similarity), and higher values means more complex signal. For example if person is actively thinking about something,there should be higher values of Sample Entropy, against someone who for example sleeps. In general the more the task is the higher value of Sample Entropy. As an advantage of Sample Entropy is the independence on the recording length.

$$SampEn(m, r, N) = -\ln[\frac{B^{m+1}(r)}{B^{m}(r)}]$$

Figure 2.27: Sample Entropy formula

Mathematical formula for Sample Entropy is in the figure 2.27. Where m is the length of EEG segment, r is tolerance and N is the total length of EEG signal. Bm̂(r) are the probabilities that the two sequences of EEG similar for m points, shall remain similar at the next point. The role of variable r is important and empirical findings suggest to use values between 0.1 to 0.25.

Approximate Entropy Another measurement of the signal complexity, lower values of Approximate Entropy indicates sinusoidal behavior of the signal, high values corresponds to a random walk signal. Compared to Sample Entropy, Approximate Entropy is dependent on the record length, minimum is recommended as 1000 samples. Another different between Sample Entropy and Approximate Entropy is that Sample Entropy does not count self matches. There is another problem with Approximate Entropy and that is the lack of relative consistency. This means that the method is dependent also on unwanted conditions of experiment.

In Approximate Entropy, the presence of repetitive patterns in signal make the signal more predictable than if there are no repetitive patterns. Approximate Entropy represents the likelihood of another similar observation occurring after the other similar patterns observations [59]. Therefore signal, which contain more repetitive patterns will have lower value of Approximate Entropy, on the other hand less repetitive patterns will have higher values of Approximate Entropy.

There are also some advantages why we should use Approximate Entropy, such as lower computational complexity or robustness against noise. Approximate Entropy analysis was used for example to EEG classification of schizophrenia, epilepsy or meth addiction [64].

Differential entropy Differential entropy is a concept from information theory that quantifies the uncertainty or randomness in a continuous random variable. In simpler terms, it's a measure of the 'surprise' or 'information content' you can expect when observing a value from a continuous distribution.

In the context of signals, like EEG readings, a lower differential entropy indicates a less complex signal, implying more predictability or self-similarity. Conversely, a higher differential entropy suggests a more complex signal with less predictability.

However, unlike with discrete variables, differential entropy can be negative, because the probability density for a continuous variable can exceed 1. Also, it can change with simple rescaling of the data, making interpretations tricky.

Hjorth complexity Hjorth complexity is defined as the sharpness of the tracings. Method entirely based on time domain but can be derived from the statistical moments of the power spectrum [41]. Can be understood as a frequency spread of the sinusoidal wave, therefore pure sinus yield zero complexity [37]. So the more far the signal have from sinus (more deviance) the higher value of Hjorth complexity as a result.

Mr. Hjorth defined activity, mobility and complexity as a good set for description of certain patient states. The activity is quntified by means of the variance. The mobility is defined as the square root of the ratio between the variances of the first derivative and the amplitude. Complexity parameter is dimensionless and derived as the ratio between the mobility of the first derivative of the EEG and the mobility of the EEG itself [Hjorth].

As for this thesis, the method looks useful but the similarity with the other entropy methods is significant. Most probable is that only one of them will be part of the input for final classifier.

Hurst Exponent This methods returns high values for self-similar signal, otherwise return lower values as output. The Hurst exponent is also called as „index of dependence“ or „index of long-range dependence“. It measures the tendency of a signal to either regress strongly to the mean or to cluster in a specific direction [47]. In neuroscience, this method is able to detect for example epileptic seizures, all in real-time. Hurst Exponent is also commonly used to estimate fractional or scaling property of EEG [65].

Originally the method was invented for Nile river flooding prediction, then it becomes practical in financial markets and that is due to a power of analyzing a long-term, hidden trend. This is an unique advantage about this technique.

$$H = \frac{\log(R(T)/S(T))}{\log(T)}$$

Figure 2.28: Hurst Exponent formula

Hurst Exponent measures trends persistence, randomness, or mean reversion. Values can be only between 0 to 1. Formula for calculating the Hurst Exponent is in figure 2.28. The method used for calculating Hurst Exponent is called R/S method, which uses R representing difference between maximum and minimum values of the signal over interval T, and S is the standard deviation of the signal interval T.

2.3.6 Feature Extraction: Connectivity analysis

EEG connectivity analysis is gaining popularity since 2005, as it can measure the connection between two EEG signals or more EEG signals, therefore it returns information about how are the brain parts connected while doing a specific task or even when being at a rest. This

information helps us because of its naturally discriminate features, and also thanks to side domain knowledge about various brain networks has advanced a lot in a past few years.

Brain as we know, has many different regions which are corresponding to various type of processing. Some of them are described and serves for example as vision control centre, or language center. Having such information enhance the connectivity analysis results, this is because knowing the certain non-zero connection information, going from one part1 to another part2 has something to do with the part1 observed function and part2 observed function (behavior in different states).

Experiments shown that the brain is coordinately activating various parts of the brain to complete different tasks. This coordinated activating is dynamic and changes during processing (executing) task. This activity can be measured even in rest state.

Calculating such analysis is done through computing bivariate measure between pairs of time series from regions of interest. Resulting into connectivity matrix or graph. Connectivity analysis can be further categorized into sensor-level connectivity analysis (has more disadvantages) and source-level connectivity analysis. Another categorization is into full brain analysis, or into just regions of interest analysis. Analysing just a regions of interest is significantly less complex and faster, making it more suitable for this thesis as the response of classification algorithm should not be far from real-time analysis.

Talking about categorization, another way to categorize connectivity analysis is functional version and effective version. Functional version measures strength of connection between various part, omitting the direction of connection. This means the analysed parts are affecting each other with the same share. The connectivity matrix will therefore be symmetric. Functional type of connectivity analysis is computed using correlation (time domain) or coherence (frequency domain). On the other hand, doing effective connectivity analysis brings information which contain the direction also. Again this means that if we analyze connection between two parts of the brain, there will be two values, one describing the measure of first part affecting second part, and other describing the opposite. Effective type of analysis carries richer information about course of events, but it is also more complex and slower to compute. This type of connectivity analysis is estimated using Granger Causality.

Correlation

Correlation is non-parametric, linear, non-directed method. Another characteristics are its sensitiveness to volume conduction and that it functions in time domain as it is not frequency specific. One type is cross-correlation which returns a measure of similarity of two signals as a function of displacement between them. Second type is auto-correlation which is same as cross-correlation but here only one signal is being considered, therefore it is about cross-correlation where both signals are the one same signal. Correlation coefficient provides the strength and direction of a relation between two variables. Again two most widely used coefficients are being used. One type is Pearson correlation coefficient, another is Spearman correlation coefficient. Main difference is that Spearman correlation evaluate correlation only by looking at the direction of a trend. Pearson, on the other hand, watch also the shape. Best score will get a distribution, which is perfectly linear. In the case of

Spearman, the only thing needed to get perfect correlation is the unbroken positive trend across the whole signal. Spearman correlation is more widely used when working with EEG, it has something to do with better outliers processing, also because the non-linearity nature of the brain function, Spearman correlation is more suitable.

One way to increase the amount of information from this analysis, is to pre-process the input data in a way that the signal is divided into individual signals of specific frequency bands. Doing that adds an information about frequency into our analysis, this is also called as cross-frequency coupling. Another way of enhancing gathered information is to vary the size of correlation window, this gives us a better granularity of information about time domain.

Coherence

Coherence, similar to correlation but operates in frequency domain, instead of a time domain. Again this method is linear, non-parametric and non-directed. As opposite to correlation, this method is frequency specific (works in frequency domain). So this method calculates the correlation, but in frequency domain. Coherence method can be enhanced by segmenting signal as it preserve more information about time.

Mutual information

Time domain analysis, which is nonlinear, non-parametric and non-directed. Tries to represent the amount of shared information of two random variables. The returned value is therefore the difference of joint probability and the product of marginal distributions of each of the two variables. Mutual information can also be estimated from entropies. In this case the value equals sumation of individual entropies and then substracting the joint entropy.

Granger causality

Granger causality is used to estimate the effective type of connectivity analysis. Granger causality can be understood as a statistical test, determining whether one signal interval is useful in predicting another. Granger defined causality as if variable one causes variable two, then predicting values for future variable two should give better results when using information about past values of variable one also, not only information about past values of variable two. That is why is said that Granger causality is a statistical concept of causality based on prediction specifically multivariate auto-regression. So it is linear and parametric. Another characteristics are that it is directed and in time domain only.

Partial Directed Coherence (PDC)

Study [30] definded Partial Directed Coherence as a Granger causality measure in the frequency domain, which is often used to infer the intensity of information flow over the brain from EEG data. Interesting thing is, that this method was able to distinguish different brain effective networks, from EEG signal. Study that does this kind of work [44], boosted the performace of PDC by combining with principles from graph theory.

Phase Slope Index

Interesting technique developed in 2008 [57], which goal is to estimate the direction of information flux in multivariate time series. Phase slope index is not sensitive to mixtures of independent sources. Another advantage is that it gives a meaningful results even if Phase spectrum is not linear (case of EEG) and the contributions from different frequencies (frequency bands) are properly weighted. Positive returned value of the phase slope index indicates that the first signal is leading the second signal. Nevertheless, phase slope index technique is limited by being a bivariate estimator, brain activity is by nature multidimensional.

Graph theory

Graph is an mathematical structure made up from nodes and edges. Graph theory is a study of graphs, which allows us to look deeper in the principles and effective processing of those graphs. Some well known problems are for example shortest path problem, isomorphism problem or largest complete subgraph, this is called clique problem. Two main types of graphs are directed and undirected, in the undirected graph two nodes has only one connection which represents joint connection. In directed graph there would be two individual connections, each representing how each node affect the another.

Example of usage in EEG analysis is that we consider every channel as a node, another could be that each localised source of signal is a node. Graphs have many characteristics, some of them are for example network density, in/out degree, regularity, completeness, characteristic path length, clustering coefficient and so on.

Network density corresponds to a total number of in/out degrees of all nodes in a network. The in/out degree is the number of connections going in and out of an node. Regularity means that every node has the same number of connections, in case of directed graph the condition is stronger because the number of degree in connections must equal number of degree out connections for each node. If graph is complete, then every pair of nodes is connected by unique connection. Characteristic path length represents the average distances between all pairs of nodes, it can be understood as how well the network is integrated or how easily can information flow through network. Clustering coefficient ranges between 0 and 1 and represents the ratio of actual connections and the maximum number of those connections. If the clustering coefficient is near to 1, that means that network is well interconnected and loose of some node should not affect this network more than network with lower clustering coefficient. Therefore it contains information about some sort of redundancy in a network.

Recent studies in a field of neural networks proposes the use of graph neural networks. Those networks applied on suitable data, can capture much more internal information about connectivity. Using a graph neural network, could set weights optimally to loss function and function as a base knowledge part of the final model. GNNs are designed to handle data represented in graph form and have demonstrated exceptional performance in various domains, including social network analysis, recommendation systems, bioinformatics, and more.

2.3.7 Feature Extraction: Microstates

Microstates analysis is an very powerful tool with many possible usage variations. The microstate analysis first step is to define how many microstates we will be distinguishing, often the choosen number of different microstates is 4. Then we calculate the global field power (GFP) peaks. These GFP peaks are then spatial clustered and mapped on one of the microstates. So the input is the original signal and output is this signal represented as a sequence of microstates.

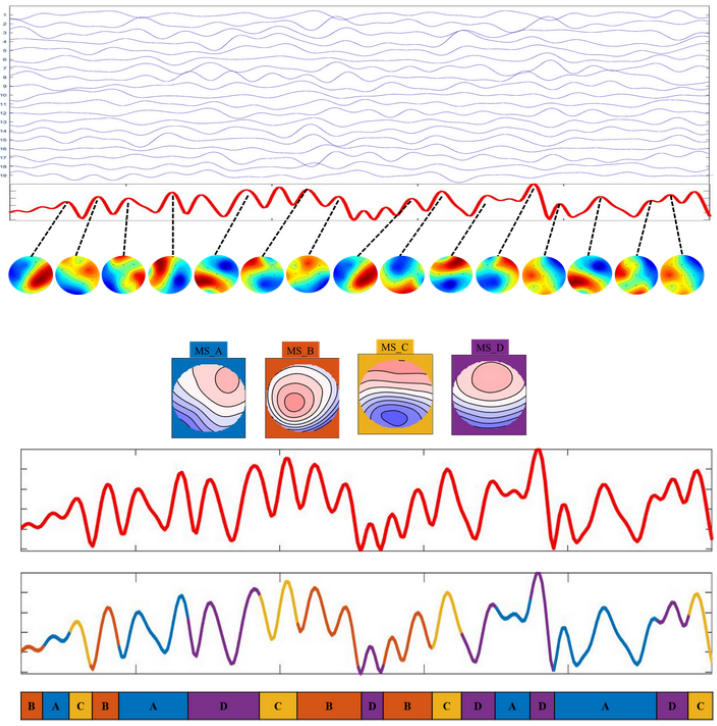

Figure 2.29: Microstate gathering process

Features that can be extracted from the microstate sequence could be for example the total occurrence of each microstate, or the duration of each microstate. Another feature could be the contribution of each microstate, which is calculated as duration multiplied by occurrence. The count of transitions between microstates could also be a valuable information, extending this will lead to co-occurrence matrix representing how often microstate1 occur after microstate2 and otherwise, or even how often microstate1 occur after microstate1. Recurrence between microstates and between transition is also a possible feature set.

2.4 Machine learning

This section is trying to discuss he relevant machine learning techniques, that may be useful when classifying emotions from EEG signal. Closer look will be taken at K-nearest neighbor classifier, random forest and Support Vector Machine (SVM), as those traditional machine learning methods are the most widely used when dealing with such a classification problem.

Further in this section, the more complex deep learning will be discussed. In deep learning subsection most of the attention will be dedicated into Convolutional neural networks, then recurrent networks and finally classical feed forward neural networks. Only the order will be reversed.

2.4.1 Traditional methods

Traditional methods are those methods which may use neurons and stack them into a layer, but the number of hidden layer stays one. Other methods can function even without creating any layer or neuron. This is for example k-nearest neighbors method.

K-nearest neighbors

This method is an non-parametric, supervised and statistics based method used for regression and classification problems [58], which principle is in classifying new data based on distance to other data. For example if new data is being classified, the classifier looks at the k nearest already classified examples. One of the advantages of this method is that it does not need training, but this carry also it's disadvantage, and that is that the method needs to remember all examples, in other words this method needs to remember all the neighbors to know how to classify. Method another advantages are that it adapts easily and require only two hyperparameters, that are number of searched neighbors „k“ and a distance metric used. But there are also significant disadvantages. One disadvantage was already mentioned, and that is that the method has to remember and therefore it does not scale well. Another disadvantage is the curse of dimensionality, which means that the method is loosing the power for classification as the dimensionality increase. This problem contributes o another problem that is he method is prone to overfit.

Figure 2.30: K-nearest neighbour function [11]

There is again a lot of variations for this method, it may be categorized based on used distance measure, most used metric is euclidean which is intuitive ad easy to understand, another metrics could be Hamming, Minkowski or Manhattan distance. Another classification is into soft k-nearest neighbors and hard. Soft rather that directly saying that the example belongs into class Z, it rather express the belonging to a class as a probability.

For our purpose this method will be used mainly in the process of gathering important features, rather than as a classifier as it emits low power in distinguishing classes of emo-

tions. Data are in addition a high dimensional problem, which this method is not good at classifying. Nevertheless, we will use this method for classifying, it is because the study [28] shows some results. This will be explored in future experimentation.

Random forest

Random forest are sophisticated method for taking decision based on more than one decision tree [42]. So it is an ensemble approach to do regression or classification. One of the advantage of this method is that it is able to take a large number of features, as one decision tree take only a subset of them. This is also the reason why random forest is trained quickly compared to other classifiers [27].

Support Vector Machines

Is another supervised learning method, which is used for regression or classification. It is able to solve linear and also nonlinear problems. Main goal of this method is to maximize the separation boundaries. Nonlinearity is solved by kernel functions, which are mapping the input data into new higher-dimension where the linear separation is possible (linear multidimensional hyperplane is a curved line back in just 2D). Correct selection of those kernel methods is very important. Some of the most well known and used are linear, polynomial or Gaussian. Overfitting of this method is partly solved by customizing and selecting the proper margin. Support vector machines are used in dozens of studies dealing with different types of analysis of EEG [54].

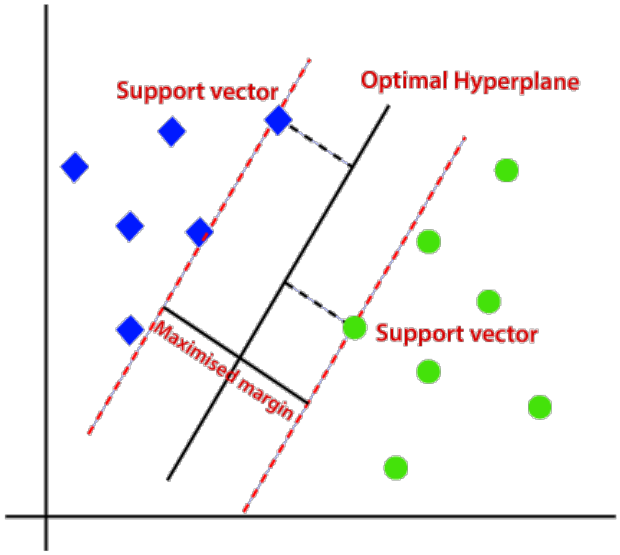

Figure 2.31: Support Vector Machines function

2.4.2 Deep Learning

Deep learning methods are currently the most widely used, and also considered as most powerful methods for doing such analysis. If for example we take a look at any competition dealing with classification problems, leading solutions are based on using some sort of this deep learning phenomena. Deep learning methods are functioning differently mainly

because of the higher number of hidden layers of neurons. More layers allow more complex problem to be solved. For example if the problem is too complex the simple neural network would not be enough, therefore increasing the number of layer also increase the overall power of the classifier. Another characteristics is the type and number of used neurons. Choosing the right type of the neural network as such will be discussed in next subsection.

Forward network

Most widely used version of deep neural network is currently classical deep forward network, which allows for effective computing as opposite to fully connected networks. Forward networks are often build in dense way, that means that every neuron is connected to every neuron in the next layer. Very deep forward networks are able to learn to distinguish very complex pattern, but it's size and complexity (of the neural network) can be also it's disadvantage. For example if we consider dataset where the complexity of classification is less complex, the classifier would be keen to overfit over time. Overfitting is one of the major problem pertaining to learning a deep neural network. Solution for overfitting in case of neuron networks is often in establishing some restrictions. Common are L1, L2 norm or dropout. Another disadvantage of using just forward network is that the classifier is not that powerful when dealing with time series. This is a problem, because our data (EEG), are full of important information seating just right in the transitions over individual records. This issue is limitating in case of classifying EEG recording. One option is to use different type of network (one that is more suitable to work with a time series data) or to prepare inputs in the way information about transition is already encapsulated inside.

Recurrent neural network

Another type of neural network is the recurrent one. Recurrent neural network address to solve the lack of power of classical feed-forward network when dealing with time series data. Recurrent neural network shortly RNN is similar to forward type of network, with the cyclically connected layers. Network is using the temporal correlations between the data at different time points, which in standard neural network is considered independently. In recurrent nets the cyclic inter-layer connections provide knowledge the way to fluctuate through the network. Cyclic connections or in other word feedback loop in network allows to capture the dynamics of the time series data, in our case this should lead to much more powerful tool for analysis, if using features without time information encapsuled. Two basic approaches are used for learning recurrent networks, Real Time Recurrent Learning (RTRL) [55] and Backpropagation Through Time (BPTT) [39]. Even the recurrent neural networks does have the drawbacks, main two of them is the exploding gradient and vanishing gradient. These problems causing difficult to propagate through extended time interval [48] [61]. Imagine the longer text which is telling the story, recurrent neural net is able to detect the relationships of words appearing near together, for example „going to gym“ will be well grasped, but the information from start of the story has a difficulty to propagate into end. In other words when recurrent neural net reach the end of the longer text, it hardly takes the information from start into a count. This problem is crucial and can be solved by using ensemble of methods targeting the longer temporal information, using heuristics or two methods was delivered in recent years. Two mentioned methods are the models called Gated Recurrent Unit (GRU) and Long-Short Time Memory (LSTM) [14]. Another way how to deal with vanishing gradient are residual connections (skip connections), which ensures better gradient propagation.

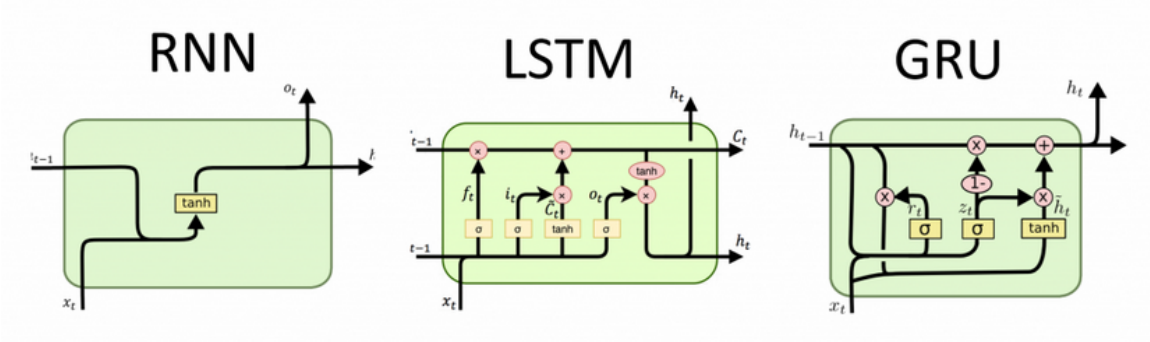

Figure 2.32: RNN x LSTM x GRU schemes [15]

Almost all of the recent state of the art solutions use LSTM or GRU, despite of much higher computational complexity, results of these evolved variants of classical RNN are just much better in major of cases.

Convolutional neuronal networks

Convolutional neural network is a type of neural network which is able to find general patterns in variety of data. The first convolutional like neural network was neocognitron [38] and its goal was to classify hand written numbers. Today, the convolutional neural networks are able to distinguish between thousands of classes, for example cat, dog, pianist, avocado and so on. Principle behind these networks is to gradually extract features which are abstract, general and hopefully discriminative. For example using convolutional network to classify between cats and dogs, the convolutional network learns what is typical to be on the picture if there is a cat, and vice versa for dogs. In the case of cat, the network should find patterns like pointed ears, these patterns should be discriminative as mentioned above. Another important feature of those patterns is that the ear can be found in any position, not only when the cat has the ear in the right top corner of the picture.

This type of network is also used as just a feature extractor, which is another way how to empower the final classification accuracy. The way of thinking about these features could be similar to that with cat ears. Every signal can be an input for a convolutional neural network, and each network should output different features. It is therefore very important to correctly design its architecture and other characteristics.

The convolutional neural network is created from 4 main components. Those components are convolutional layers, activation layers, pooling layers and the fully connected layers at the end. The convolutional layers are applying the convolutional operation on its inputs. Important characteristics of this layer is the step S, and padding P. The importance of these two parameters is significant as it decide what dimension the result will be. An example of the consequences of different values for step S and padding P is shown in pictures 2.33 and 2.34. Another characteristic of an convolutional layer is that it consists of a many filters which are all applied of the input data, and all should lead to different results. Activation layers are often the non-linear part of network, in convolutional neural networks, Rectified Linear Unit (ReLU) or hyperbolic tangens. ReLU has a certain advantage in its computational efficiency. The reason why ReLU is computationaly efficient is that values below zero are mapped onto zero and therefore they are omitted in further

computation. The pooling layers are dimension reducing mechanism. Pooling layer must use some strategy for doing the dimensionality reduction, which most often is either max pooling or average pooling. Max pooling takes for example 2x2 input and keep only the highest value, average pooling does the average over all and then keep this average. The last component, the fully-connected layers are the same layers as in classical forward networks described above. These fully connected layer uses almost exclusively the softmax activation function.

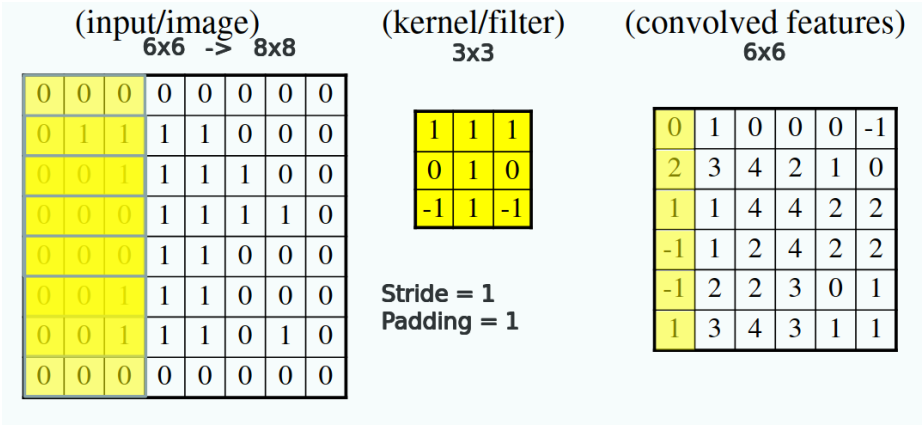

Figure 2.33: Convolution with S=1 P=1

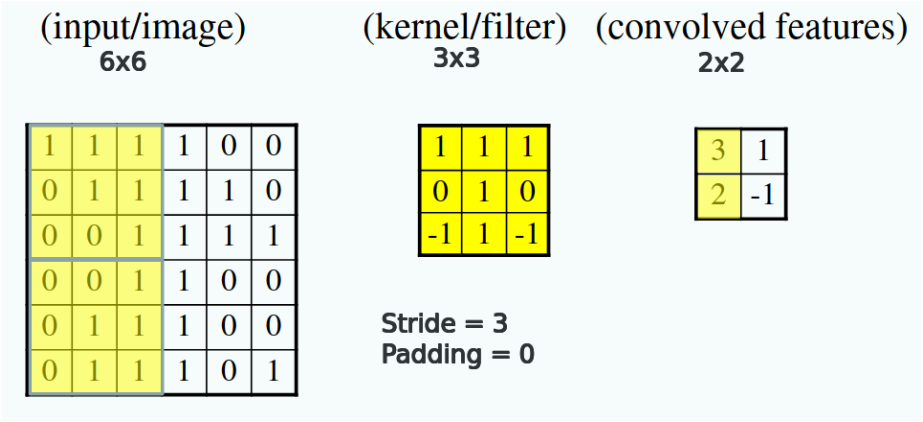

Figure 2.34: Convolution with S=3 P=0

State of the art architectures are a big models consisting of plenty of neurons, utilizing number of heuristics like dropout technique [67], or the 1x1 convolutions which speeds up the computation if used correctly. Another often used techniques are the Local Response Normalization Layers and Flattening layers. One of the problems of Convolutional neural network is pretty similar to that one from recurrent neural networks, that is vanishing gradient. Inception block from GoogLeNet [**?**] or building block from ResNet [40] are trying to solve this issue by various strategies. Main method used to train Convolutional neural networks is Mini-Batch Gradient Descent, which uses the backpropagation principle. Backpropagation behaves according to the type of layer it is going through.

EEG emotion recognition is an multidimensional, non-linear and nonstationary problem which can be solved thanks to Convolutional Networks, as can be seen on many competitions solving the similar task. For example competition Grasp-and-Lift on Kaggle website [12], where the winners were using Convolutional neural networks intensively. The second best solution which has accuracy over 98% used ensemble of almost only the Convolutional neural networks. Other competitors tend to use other features also. The final model uses the ensemble of all the chosen methods and features, therefore final processing pipeline is implemented via decently complex architecture.

Graph neural networks

The use of graph-based deep learning models is a relatively novel approach to EEG data analysis, specifically for the task of emotion recognition. Graph-based models are particularly advantageous because of their ability to capture the rich topological structure present in EEG data, which often cannot be effectively utilized by traditional deep learning models such as feed-forward neural networks (FNNs), recurrent neural networks (RNNs), or even convolutional neural networks (CNNs).

At a fundamental level, a graph is a mathematical structure consisting of nodes and edges. In the context of EEG data, these nodes often represent channels or electrodes (SEED dataset - 62 channels), while the edges represent the functional or structural relationships between them (this is the main additional knowledge utilized against common architectures). The human brain is essentially a complex network, with neurons and their synaptic connections forming a natural graph-like structure. Therefore, representing EEG data as a graph can intuitively mirror the inherent connectivity structure of the brain, thereby allowing us to extract and exploit information that might otherwise be missed.

GNNs are a class of deep learning models designed specifically for graph data. They learn to aggregate information from a node's neighbors in the graph, which can effectively capture the local and global structural information in the data. For example, Graph Convolutional Networks (GCNs) generalize the operation of convolution to graph data, enabling the extraction of local features in the node's neighborhood.

Because in EEG-based emotion recognition, graph-based models can be particularly beneficial. Emotions are thought to arise from the dynamic interplay of different brain regions. Therefore, understanding the functional connectivity between different EEG channels can provide valuable insights into emotional states. Graph-based models can capture these interactions and leverage them to make more accurate predictions.

For example, in a recent study, a graph theory-based approach was used to construct functional brain networks from EEG data, and graph measures like clustering coefficient, characteristic path length, and modularity were used to characterize these networks. It was found that these measures could effectively differentiate between different emotional states.

2.5 Performance tuning, Metrics

This chapter presents an in-depth investigation of performance tuning through cross-validation and a thorough evaluation of classification metrics for the developed model. Cross-validation serves as an tool for validating the model's performance and adjusting its hyperparameters,

thereby enhancing its generalization ability. The study further employs a range of metrics, including accuracy, confusion matrix, precision, recall, F1 score, and Area Under the Curve (AUC), to assess the model's predictive prowess. These metrics illuminate different facets of the model's performance, informing the direction of subsequent optimization efforts.

2.5.1 Cross Validation

Cross Validation is a method for evaluating statistical analysis generalization performance. Goal of this method is to optimize the hyper-parameters of the classifier, for example number of hidden layers or number of neurons. The principle behind is that every time someone is training and then testing the supervised predictive model, it is done first on the training part of the dataset (this is usually around 70% - 90% of all data) and then on the rest, which is used for the testing part. The performance of the model on the testing part is sometimes called generalization performance, it means how good is the model when predicting on unseen data. Problems associated with are overfitting to training data and selective bias. Overfitting is a well known problem that represents very good accuracy when predicting on the training data, but very low accuracy on the testing (unseen data) part. Selective bias targets the problem when dataset contains records from different sources. EEG datasets are exactly one of those datasets burdened with selective bias problem as the recording is almost every time done on more than one participant.

In order to have more control over training the model, instead of using just training and testing split, data is split into three parts, training, validation and testing part. Validation part of the dataset will be used for testing the performance in every epoch. The way of choosing the validation part is one of the aims of the cross validation methods. Cross validation uses different validation sets over all epochs.

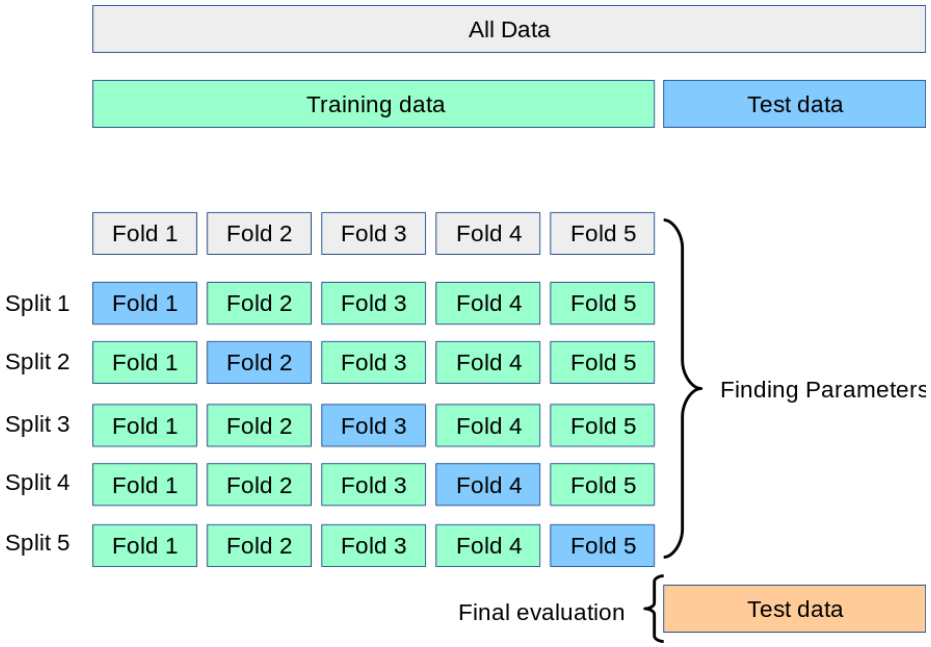

Figure 2.35: Grid search cross validation

There are 2 main types of cross validation, exhaustive and non-exhaustive. The first group contain approaches that will try to train and test on all possible ways to divide the original dataset. Those are leave-p-out cross validation and the special version leave-one-out cross validation. Leave-p-out cross validation takes p observations as a testing set in one epoch of training, than it moves to next p observations and select them to be new testing data, the old p observations will be considered as training data in next run. Leave-one-out is same but p equals one. The non-exhaustive approaches are the approximation of exhaustive version as it doesn't try every possible ways of dividing the data.

Cross validation should lead to better performance and also it should partly solve the overfitting and selection bias problems. Another problem that is partly solved is the information leaks from validation dataset to model in process of training. It is important to mention that shuffling the data could be necessary in some cases, but also could lead to other problems. In ordered dataset of numbers, if split is done, testing dataset could contain numbers which are not present in the training dataset, this leads to meaning less training. On the other hand shuffling the time-series data (EEG recordings) could lead to predicting past from future information, which is not desired. Overall, if cross validation used carefully and in proper way, then it returns valuable results.

2.5.2 Performance metrics

First metric that comes to mind is the accuracy of the model prediction. This is one of the metrics that is right to use but it delivers much less information about performance than for example confusion matrix score or ROC curve.

Accuracy

Accuracy of the model is simply the percentage of correctly classified classified examples. But imagine that we have to classify when there is a cat or dog on the picture, then if we have data containing 99% of dogs, predicting dog every time gives us accuracy 99%, but the model doesn't have any intelligence in it, hence accuracy is not a very good performance metric.

Confusion Matrix

Confusion matrix represents better way of evaluating the model performance, since it divide the results into classes forming an NxN matrix where N is the number of different classes. Then weak part of classifier can be easily seen. For example if we have the numbers as a picture and we have to classify them, in confusion matrix we should observe that the model tend to classify 9 or 6 as 8, since it is similar in shape. On the other hand classifier will most probably classify 7 correctly as its shape is most different from all the other numbers. This is easily readable from the confusion matrix. Confusion matrix can be used only with discrete space, so it can be used on valence-arousal model only after categorization.

Receiver operating characteristic curve (ROC)

This metric tries to plot results from classifier into 2D plane, in which one axis represents recall (y axis) in other words True-positive rate and the other axis False-positive rate.

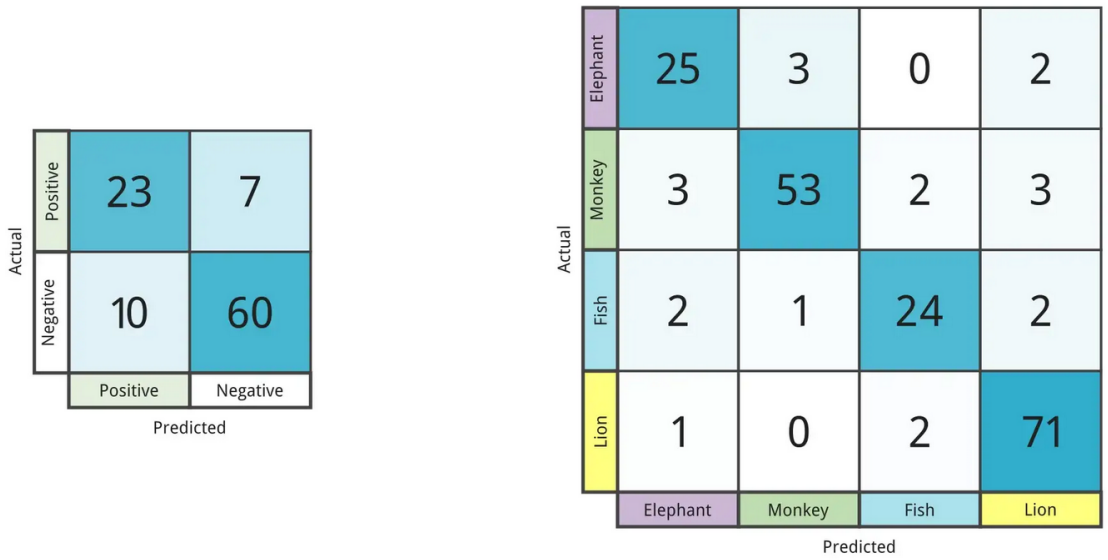

Figure 2.36: Confusion matrix [5]

The goal of this metric is to show model performance in different point of view. Ideal performance is the leftmost and topmost point. But it is also important to choose, based on the problem we are solving, some problems appreciate false alarm over miss, for example when predicting tumors, it is better to check even if the probability is small.

Since task of this thesis is multi class task, I need to use modification of original ROC. There are 4 ways of doing this:

One-vs-all Most common approach. For each class, it creates a ROC curve, treating that class as the positive class and all other classes as the negative class. This approach can be further used if patient suffers from depression, it is important to detect sadness, maybe more than excitement, that is just an example.

Multi-class ROC Area This is an extension of the ROC for multi-class problems. It calculates the pairwise class ROC area and averages them.

Micro-average In micro-averaging, all the instances/classes contribute equally to the final average. It aggregates the contributions of all classes to compute the average metric. In a multi-class classification setup, micro-average is preferable if you suspect there might be class imbalance (i.e., you have much more instances in one class than the other classes).

Macro-average The score for each class is calculated separately and then averaged. This does not take class imbalance into account. So, each class has the same weight regardless of its size. Macro-averaging is used when you want to know how the system performs overall across the sets of data.

If the macro-average is much lower than the micro-average then it means that the classifier is not performing well on the minority classes (assuming that minority classes refers to classes with fewer total instances).

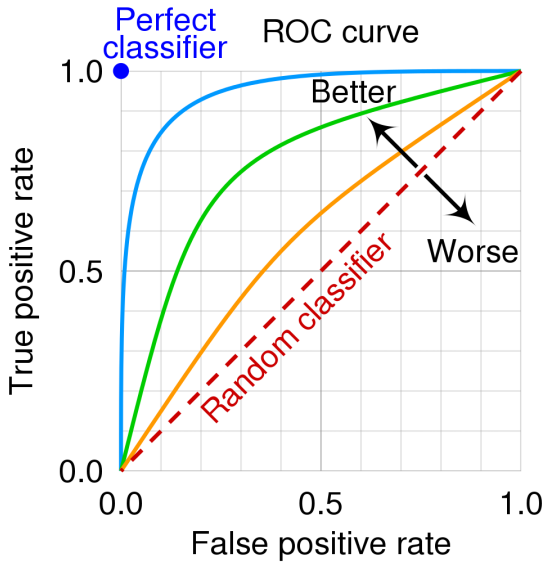

Figure 2.37: ROC curve [18]

F1 Score

Performance metric used in binary classification tasks to assess the balance between precision and recall. It is particularly useful when the classes are imbalanced, meaning that one class has a significantly larger number of instances than the other.

Precision Precision is the ratio of true positive (TP) predictions to the total number of positive predictions (both true positives and false positives). It measures how many of the predicted positive instances are actually correct.

Recall Recall is the ratio of true positive (TP) predictions to the total number of actual positive instances (both true positives and false negatives). It measures how well the classifier identifies positive instances correctly.

The F1 score ranges from 0 to 1, with 1 being the best possible value, indicating perfect precision and recall. A high F1 score suggests a balanced performance between precision and recall, which is desirable in scenarios where both false positives and false negatives are equally important to consider.

2.6 Datasets

Doing classification of emotional states from electroencephalogram (EEG) needs a high quality data in order to get good results. One dataset is usually recorded with more than one participant and also there is usually more than one recording of desired response. Moreover another important characteristics of various EEG datasets are the montage and sampling frequency. Montage describes how and how much channels were used when doing

the experiment. Common values are for example 19 channels for 10-20 montage system. Sampling frequency is usually 256, 512 or ever 1000 Hertz, the higher value means better time precision. Dataset used in this thesis are the SEED dataset [73].

2.6.1 SEED

The SEED dataset offers a rich repository of EEG data collected from 15 individual subjects. Each participant was subjected to a series of 15 different video clips, which they were asked to watch in three separate sessions. This protocol was designed to capture a spectrum of brain activity and emotional responses.

The data acquisition process was carried out at a sampling rate of 200Hz, a frequency that ensures enough capture of EEG signals over time. The EEG data, recorded from 62 channels, provides a spatial view of brain activity. The relatively high number of channels contributes to the complexity and high dimensionality of the dataset but also opens up opportunities for exploring important spatial patterns of brain signals.

Additionally, the SEED dataset is multimodal, meaning it incorporates different types of data or data representations. This characteristic enriches the dataset and allows for more complex and robust analyses.

2.6.2 SEED IV and V

Building upon the foundation laid by the original SEED dataset, the SEED-IV and SEED-V datasets represent the next stages in this series, expanding the scope and complexity of the emotional categories.

The SEED-IV dataset includes data related to four distinct emotional categories: happiness, sadness, fear, and a neutral state. This addition of fear as a category allows the exploration of brain responses to a broader range of emotional stimuli, including those that trigger a sense of threat or danger.

The SEED-V dataset further extends this emotional palette by introducing disgust as an additional category. Now encompassing five categories: happiness, sadness, fear, disgust, and a neutral state, the SEED-V dataset provides an even more diverse collection of emotional states. The inclusion of disgust allows for the investigation of more complex or nuanced emotional responses, possibly associated with revulsion or moral judgement.

SOTA The State of the art solutions for recognizing emotions from EEG, are almost every time based on some deep neural network. Still the technique is quite different. In the table 2.38, some of the proposed models are shown. Clearly, it can be seen that since SVM doesn't have the computational power equal to deep neural networks it reaches only a low accuracy score. What cannot be read from the picture is that the data preprocessing is very important in this scenario. The results shown are all of the patient dependent type, means that the learning and prediction was done independently for each patient.

The following table 2.39 outlines the state-of-the-art results achieved, now on the original SEED dataset, where are only 3 classes.

A noteworthy achievement is the performance of the Support Vector Machine (SVM) model, which achieved an accuracy of 83.99%. Considering SVM's nature as a linear classifier, this result is high. Beyond SVM, several more advanced models have been employed,

State of the art model (SEED IV dataset)

Model	Accuracy
SVM (Support vector machine)	31.46%
Dynamical Graph Convolutional Neural Networks	52.82%
NN Model based on Cerebral Hemispheric Asymmetry	65.59%
Graph-Based Multi-Task Self-Supervised Learning (unsupervised)	65.61%
Critical frequency bands and channels vie deep NN.	66.77%
Bi-hemispheric Discrepancy Model	69.03%
Regularized Graph Neural Networks	73.84%
Saliency Fusion	74.42%
Graph-Based Multi-Task Self-Supervised Learning (supervised)	86.37%
Deep Canonical correlation analysis	87.45%

Figure 2.38:

leading to progressively higher accuracies. For instance, the Dynamic Graph Convolutional Neural Network (DGCNN)[66] model achieved an accuracy of 90.4%, demonstrating the potential of graph-based neural networks in capturing the spatial structure of EEG data.

Subsequent models, including the Domain Adversarial Neural Network (DANN) [46] and the Bi-hemispheric DANN (BiDANN) [51], improved on these results, achieving accuracies of 91.36% and 92.38% respectively. These models, which utilize domain adaptation techniques, showcase the benefits of leveraging shared information across different subjects or sessions in the dataset.

Further advancements were made with the implementation of the Bi-Hemispheric Discrepancy Model (BiHDM) [50] and the Regularized Graph Neural Network (RGNN) [74], which achieved accuracies of 93.12% and 94.24% respectively. These models incorporate sophisticated architectures designed to capture complex temporal and spatial patterns in the data.

Finally, the model used in this work, referred to as the GMSS [49], achieved the highest accuracy to date, reaching an impressive 96.48%. This result is significant in the analysis of the SEED dataset and it demonstrates the power of the chosen model in deciphering intricate patterns within EEG data and accurately predicting corresponding emotional states.

State of the art model (SEED dataset)

Model	Accuracy
SVM	83.99%
DGCNN	90.4%
DANN	91.36%
BiDANN	92.38%
BiHDM	93.12%
RGNN	94.24%
GMSS	96.48%

Figure 2.39:

Chapter 3

Emotion recognition proposed method

This chapter summarizes the previous analyses, and tries to find the optimal solution idea. First there is a section about classical machine learning classification pipeline, which show the basic and un-detailed way how to solve classification problem via machine learning. Then next chapters are talking about my proposed solution for the problem of emotion recognition.

3.1 Classical machine learning pipeline

In the classical machine learning pipeline, there are some steps which cannot be ommited. First we got the data collection, which is already done in this case. As this work uses the datasets SEED and its variants, there is no more need for data collection. The data preparation, an essential step, in this case, it would be composed of denoising, filtering and normalization. The preprocessing for EEG data has its own rules (muscle artifacts, hearth artifacts). Next there is a model training, it is where the model learn. Moving to the next step there is a model checking and validation and deploying, which are the finishing moves when doing such a classifier.

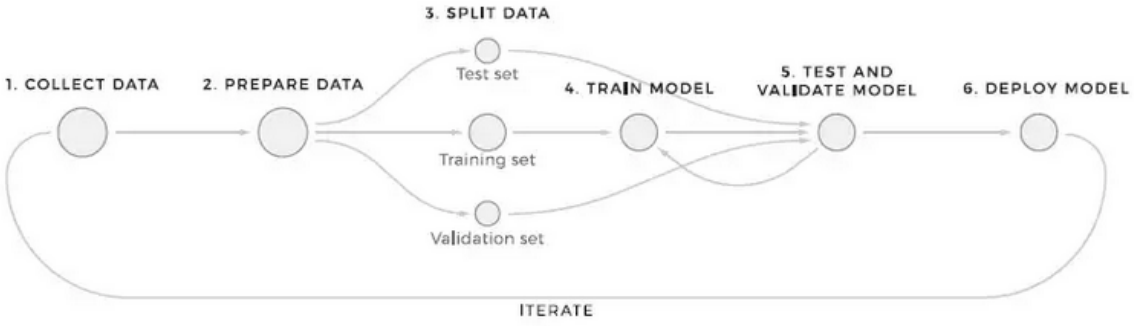

Figure 3.1: Classical machine learning pipeline [4]

As previously said, classical machine learning pipeline describes the basic process behind machine learning in the most simple way. Feature extraction will come in words of different types of analyses, like frequency, time, source, connectivity or microstates.

Our data, will yet need to be cleaned first. Or if any inconsistencies happen, those must be figured out before moving to artifact cleaning. After that cleaning process via Independent Component Analysis, and then Canonical correlation analysis take a place. So now with cleaned data we can do normalization which sets the same importance to all channels. Important step which contributes to meaningfulness of whole analysis.

3.2 Feature Selection

In my exploration of EEG data analysis, I've embarked on a quest to identify the most informative features. The journey was not without its challenges, but it was an process that helped me refine the approach and improve the model's performance.

Initial stages were focused on frequency and time analysis, considering a wide array of methods, including entropy measurements, co-occurrences, and fractal dimensions. Additionally, the impact of different channel combinations was investigated. However, as understanding deepened, I realized that while these methods provided valuable insights, they are also redundant to each other and they may introduce a higher level of complexity that may not necessarily enhance our final classification results.

Similarly, I have considered source localization and connectivity analysis, recognizing that these methods could potentially unearth unique types of information. For instance, we hypothesized that by assigning more weight to the specific brain regions could enhance our model's sensitivity to emotional cues. Yet, we also acknowledged the potential risks associated with this approach, such as the possible loss of important information due to the complex dynamics of brain networks. Or as the source localization is highly computationally demanding, I have tried to figure out different way of capturing such information.

After much deliberation and experimentation, we decided to adopt a more tricky approach. I chose to forego the source and connectivity analysis, focusing instead on building a self-supervised techniques, that hopefully capture the desired source and connectivity insights of EEG.

Numerous feature combinations were systematically examined with the goal of identifying the most informative subset. Each iteration deepened my understanding of the nature of EEG data and its correlation with various emotional states. Ultimately, the feature set chosen for the models was the combination of differential entropy and power in frequency bands. This feature set was selected as it captures both time and frequency information, while the self-supervised tasks encapsulate source and connectivity information. This comprehensive feature set is expected to yield promising results in the final model performance.

3.3 Machine learning

Machine learning and especially deep machine learning is the grail and on of the most powerful technique here. As first the Convolutional neural network will be used in high variety of ways. Convolutional neural networks have the power to extract features in different levels of abstraction. Some of the most accurate models competing on the kaggle website, used only those. My aim here is to use them as feature extractors.

Classical forward dense pass neural nets are going to be used as last few layers, just to take the final features and make a prediction. So this brings yet another dose of complexity when finding the best possible model. Because of that, some strategy for exploring the space of solutions should be considered. This work is going to use only variations of three neural nets types, forward, recurrent and convolutional. Other types also exists for example Boltzmann machine, Self Organizing Maps, Learning Vector Quantization etc., but they are not considered in this work.

As said before, third used type is going to be the recurrent nets. The recurrentness gives us ability to better understand the sequential data, because there is a feedback loop, the information from past flows through recurrent network processing to the present. Recurrent networks suffers from the vanishing gradient phenomena, which is partly fixed by using more intelligent, also complex, units. Units are the LSTM (Long Short Time Memory), and GRU (Gated Recurrent Unit). These are harder to train, but should reach better accuracy.

The primary objective of this part is to build an system composed of more parts. First part, probably the convolutional networks able to output meaningful features (further being mixed with standard methods). Second, there will probably be recurrent part, which is going to be GRU or LSTM. The recurrent part takes features from convolutional neural networks and other methods previously mentioned. Output of the recurrent part will be directly connected to the last part, which is a dense forward network. The goal of the last part is to take the outputs from recurrent part as an input and do a final computation which results in desired final outputs.

3.4 Alternative ways

In order to set a good trade off between performance and speed or cost, the alternative computation pipeline will also be considered. The principle behind this technique is to figure out where the information gain is not good enough to go through next computation block, because of lowering the computation time and cost. Therefore from the original plan pathway in form of CNN->RNN->FNN, in some cases the RNN or CNN could be omitted. The granularity of this solution would reach much greater softness. The goal of this is to retain in almost the same performance score while lowering the complexity. Omitting the genetic algorithms in advanced phase will also be considered. Gathering necessary information will be done by experimentation.

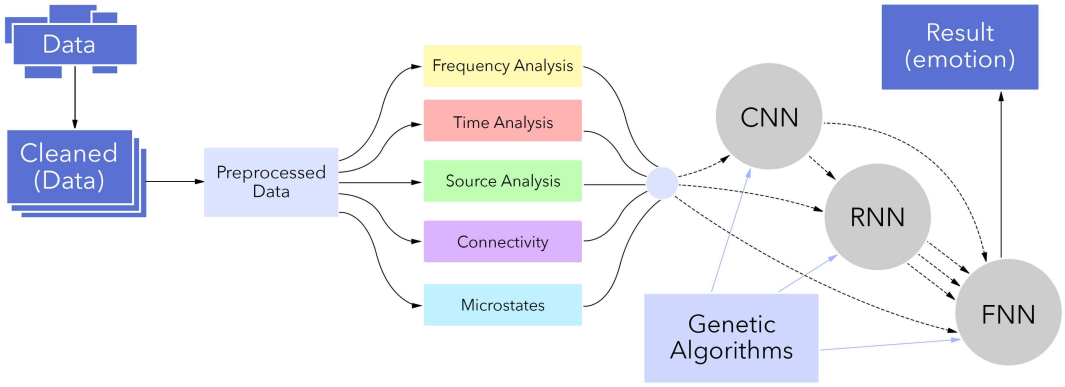

Figure 3.2: Proposed processing pipeline

Performance The performance evaluation will be done through few measures. Accuracy will be first, second is confusion matrix, third will be precision and recall, fourth is micro ROC, and the last one the F-1 score. F1-score can be calculated as two times (precision×recall)/(precision+recall).

Chapter 4

Implementation

4.1 Feature Selection

4.1.1 Frequency features

As said before, EEG signals are usually decomposed into frequency bands (delta, theta, alpha, beta, gamma), and these bands have been linked with different emotional states in numerous studies. Thanks to decomposition to frequency bands, we get better resolution and hence more precise inputs. Now the question is which analysis brings most information.

Features from analysis of amplitude and asymmetry achieved only a poor score on SEED (3 classes). It could be suffering form lack of information contained in both amplitude or asymmetry, or the information requires more complex function. Overall, the feature that represents frequency domain information is represented by relative power. As relative power is better for learning neural network than absolute. Relative power is the power in a specific frequency band relative to the total power across all frequency bands. It is often used to account for individual differences in overall EEG power. The power in different frequency bands is a common and well known informative feature used in emotion recognition. For example, an increase in alpha power (especially in the right frontal region) has been associated with increased positive emotion, while increased beta power has been associated with increased negative emotion or mental stress. Just to clarify, these are the observed states, when the power in specific frequency domains dominate:

Delta (0.5-4 Hz) is the lowest frequency range, which is often associated with deep, dreamless sleep and regenerative healing. Theta (4-8 Hz) range is often associated with light sleep, meditation, and the state of mind where you can access deep, profound thoughts. Alpha (8-12 Hz) frequency range is associated with relaxation, calmness, and peacefulness. Beta (12-30 Hz) frequency band is associated with normal waking consciousness, attention, concentration, and cognitive processing. Gamma (30-100 Hz) fastest frequency range is associated with high level cognitive tasks, problem-solving, and perception.

By considering the energy within these bands, this is essentially quantifying the amount of 'activity' within these frequency ranges, which can serve as an indicator of the individual's emotional state. Moreover, based on my own shallow machine learning tests using random forests, SVM, and KNN, I found that these features consistently provided good classification performance.

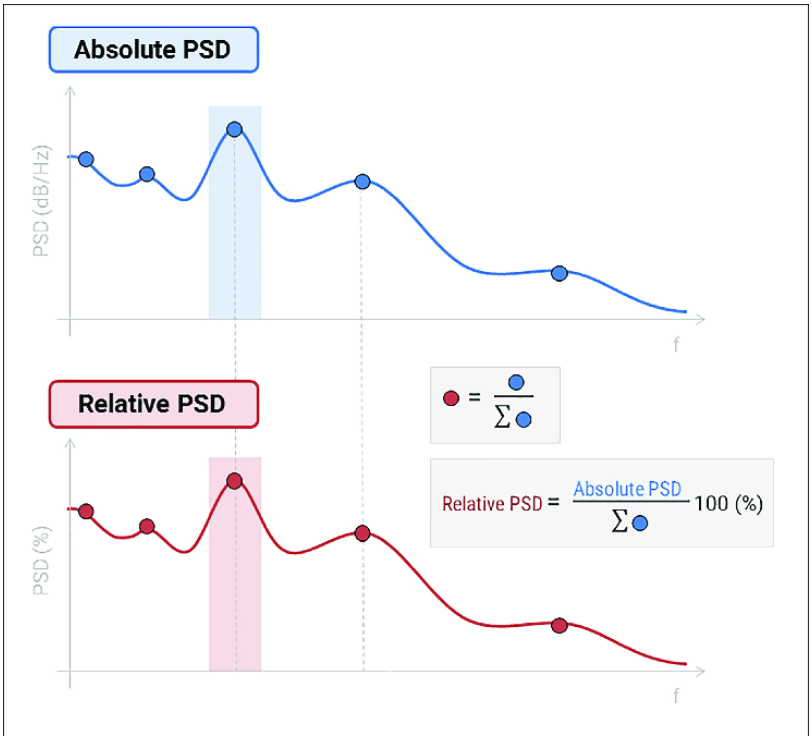

Figure 4.1: Absolute and relative power [49]

4.1.2 Time features

To get better information about time domain of given signal, I have tried many options. Hjorth parameters, Hurst exponent and different entropies. Chosen time features as for setting the weights based on them. Sample Entropy, Approximate Entropy and Differential Entropy, these are measures of the complexity or unpredictability of a time series. And Differential entropy was selected as a features for the classification. It is easy to calculate and pretty informative about time domain. Moreover this feature is used in many studies, so it is a good anchor for comparison between different solutions.

Sample and relative entropy, both of these measures can be used to analyze EEG signals for emotion recognition. More complex emotional states may be associated with more complex or unpredictable EEG signals. For example, a study might find that sample entropy or approximate entropy is higher during states of emotional arousal compared to calm states.

And now the Differential entropy. It is a measure of the randomness or unpredictability in a continuous random variable. It is similar to other entropies in term of higher complexity, higher activity. For example, a highly aroused emotional state might result in more complex EEG signals compared to a calm state. This feature can, therefore, surely provide valuable information for emotion recognition. Experiments with different machine learning algorithms (random forests, SVM, and KNN) have also shown that differential entropy is a stable, useful feature for emotion recognition.

4.1.3 Connectivity

While traditional EEG studies often use measures of connectivity, such as coherence or mutual information, to quantify the statistical relationships between different EEG channels or brain regions, in this study I decided to use a Self-supervised task specially designed for that. In addition I decided to use Graph Neural Network (GNN) too, to capture this information. This decision was done on the end of the work, so until then all models were missing such graph structure information.

The use of a GNN is motivated by the understanding that the brain is inherently a complex network, where different regions are interconnected and communicate with each other. This structure is naturally represented as a graph, where each node corresponds to an EEG channel or brain region, and the edges represent the connections between them.

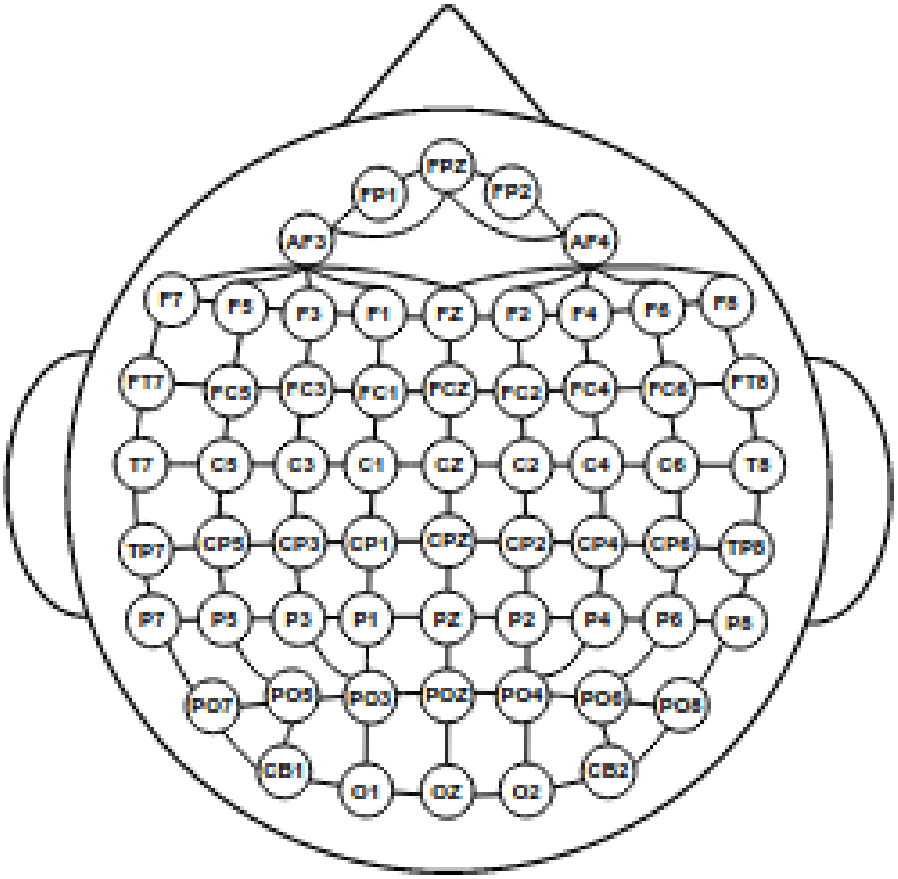

Figure 4.2: Edges in graph based neural network [49]

A GNN is well-suited to this task, as it is designed to work with graph-structured data. The GNN operates by propagating information along the edges of the graph, allowing it to

capture the dependencies between different nodes. This process mirrors the propagation of information in the brain, where activity in one region can influence activity in connected regions. This makes the GNN a powerful tool for modeling the brain's electrical activity.

4.1.4 Source analysis

In the domain of EEG analysis, source localization methods such as LORETA are often employed to infer the neural sources of the electrical activity measured at the scalp. These methods have the potential to provide valuable spatial information about the underlying brain activity. However, they are also known to be computationally intensive and require strong assumptions about the underlying brain anatomy and the number and location of neural sources, which can introduce uncertainties into the analysis.

Given these challenges, I decided to omit traditional source analysis methods in my work. Instead, I opted to use a self-supervised learning approach, which has been shown to be effective in learning useful representations from unlabelled data, and is less computationally demanding compared to source localization methods.

Later in the thesis, a detailed description of the specific self-supervised task used in this study will be described, and present results showing the effectiveness of this approach in learning useful features for emotion recognition from EEG data. My findings suggest that self-supervised learning is a promising alternative to traditional source localization methods for EEG analysis.

4.1.5 Microstates

While microstates are a rich source of information, capturing a blend of spatial, temporal, and frequency data, I have opted not to include them as features in this thesis for several reasons.

Firstly, the computation and interpretation of microstates can be complex and time-consuming. The process involves segmenting the continuous EEG data into distinct microstates, which requires careful parameter tuning and validation. Moreover, the interpretation of these microstates, especially in the context of emotion recognition, is still a topic of ongoing research, and the links between specific microstates and emotional states are not yet fully understood.

Secondly, chosen approach of using a Graph Neural Network (GNN) and a self-supervised task already captures a significant amount of spatial, temporal, and frequency information from the EEG data. The GNN is designed to capture the connectivity and interactions between different brain regions, while the self-supervised task captures source/frequency dynamics, and our chosen features capture frequency and time information. This comprehensive approach reduces the need for the additional complexity of microstate analysis.

4.2 Self-supervised learning

In this work, I opted to leverage the power of self-supervised learning, innovative strategy that has shown promise in several domains. The principal motivation behind the adoption of self-supervised learning is its capability to exploit large amounts of unlabeled data

effectively. Given that the labelling process is typically time-consuming and requires expert knowledge, self-supervised learning presents an alternative to traditional supervised learning approaches.

The self-supervised methods are widely used, for example in Natural language processing. We have this sentence „Hello, how are you?“, we can delete the word „how“. Now we have a training sample, because the ground truth is the original sentence. This idea can be smoothly transferred to EEG domain.

Now, in the process, these models learn to extract useful, different-level features from the data that can be used for downstream tasks such as classification or regression. The task designed for the self-supervised learning phase is called the „pretext task,“ while the downstream task that utilizes the learned features is called the „target task.“

In the context of EEG data for emotion recognition, I have selected to experiment with designed an appropriate pretext task that encourages the model to learn meaningful representations from the data. For instance, the model might be tasked with predicting the future activity of a given EEG channel based on its past activity, or it might be asked to reconstruct the activity of one EEG channel based on the activity of other channels. This way, the model needs to learn the underlying spatial and temporal dynamics of the EEG data, which can provide useful information about the individual's emotional state.

In my experiments, I found that this self-supervised learning approach was able to extract features that significantly improved the performance of my emotion recognition model. This suggests that the spatial, connectivity and temporal dynamics captured by the self-supervised learning task are informative of the individual's emotional state, and demonstrates the potential of self-supervised learning as a powerful tool for EEG-based emotion recognition.

4.2.1 Noise injection

As part of the self-supervised learning strategy, this study employs a unique task known as „noise injection.“ The idea behind this task is to allow the model to capture and learn the inherent structure of the EEG data by artificially injecting noise into the data.

The noise injection process begins with the original EEG data. Data consists of 62 channels, each corresponding to different brain regions. For the noise injection task, a random channel from these 62 channels is selected. This selected channel then undergoes random multiplication, introducing a form of perturbation or „noise“ into the data.

The goal of this task is to accurately classify the channel that has undergone modification. This is achieved through a specialized classification head, $\mathrm{He}(\cdot)$. $\mathrm{He}(\cdot)$, also referred to as the „noise head.“ The noise head takes as input the EEG data with the modified channel and outputs a prediction of which channel was altered.

The loss function for this task is the cross-entropy between the predicted and true labels of the altered channels. This loss function encourages the model to correctly predict the labels by learning to recognize the unique characteristics of each channel.

The noise injection task provides the model with a robust learning experience. By introducing random noise into different channels, the model is forced to learn the specific

characteristics and patterns of each channel in order to accurately predict the altered channel. This enables the model to better understand the structure of the EEG data, which is crucial for recognizing emotional expressions from the data.

4.2.2 Spatial Jigsaw

Unique self-supervised learning task is employed, known as the „spatial jigsaw puzzle“ [49] to capture the spatial patterns across different brain regions from the EEG data. This approach is predicated on the understanding that different brain regions play varying roles in emotional expression. The spatial jigsaw puzzle task thus involves creating a series of brain region permutations.

The original EEG data, denoted as X, is divided into 10 blocks according to the location of the brain regions. The goal is to determine the correct permutation of these brain region blocks. In total, there are 10 factorial, or 3,628,800 possible permutations, representing all possible rearrangements of these 10 blocks.

However, it is computationally too intensive and challenging to distinguish between such a large number of permutations. To address this, a selection operator, $\mathrm{Rk}(\cdot)$, that picks out the k permutations with the maximum Hamming distance from the complete set of permutations.

This operator is designed to select permutations that are as different from each other as possible, thereby maximizing the diversity of the training data for the self-supervised task. In this thesis, I set k to 128 (as in original paper [49]), resulting in 128 unique permutations, each with its own pseudo label ranging from 1 to 128.

Each input data sample is then randomly transformed into one of these 128 permutations, and the corresponding pseudo label is assigned. This forms the basis of the spatial jigsaw puzzle task, where the model is trained to predict the pseudo label (i.e., the permutation) of each input sample.

To recognize these spatial jigsaw puzzles, a classification head $\mathrm{Hs}(\cdot)$ (stands for spatial head) is applied. The loss function for this task is the cross-entropy between the predicted and true pseudo labels. The loss function encourages the model to correctly predict the pseudo labels, by learning to recognize the different permutations of brain regions.

Initial version The initial version of the self-supervised learning strategy in this study involved a task known as „channel shuffling.“ Unlike the „spatial jigsaw puzzle“ task, which divided the EEG data into 10 blocks based on brain regions, the channel shuffling task did not group the data in this way. Instead, it treated each channel as an independent entity, leading to a much larger space of possible permutations.

In this task, the original EEG data, denoted as X, was composed of 62 independent channels. The task involved creating a permutation of these channels, which effectively introduced a form of „noise“ into the data. Due to the larger number of channels, the total number of possible permutations was not 10 factorial (as in the spatial jigsaw puzzle task), but 62 factorial. Selecting 100 distant permutations serves as a good challenge for model.

Thanks to this spatial jigsaw puzzle task, the model captures the spatial patterns of the EEG electrodes across different brain regions, which is critical for understanding the neural basis of emotional expression. This approach forms a crucial part of our self-supervised learning strategy for emotion recognition from EEG data.

4.2.3 Frequency Jigsaw

Another form of self-supervised learning, to harness the interrelationships between different frequency bands. This task is tailored to identify crucial frequency bands for EEG emotion recognition, enhancing the model's discriminatory abilities.

Similar to the spatial jigsaw puzzle task, the objective here is to discern the correct permutation of mentioned frequency band blocks. In total, there are 5 factorial, or 120, possible permutations, each assigned a unique pseudo label.

Each input data sample is randomly transformed into one of these 120 permutations, and its corresponding pseudo label is assigned. This forms the crux of the frequency jigsaw puzzle task, where the model learns to predict the pseudo label of each input sample, thereby learning to recognize the different permutations of frequency bands.

A classification head, denoted as Hf($\cdot$), is employed to recognize these frequency jigsaw puzzles. The cross-entropy loss between the model's predicted and actual pseudo labels forms the loss function for this task. This function nudges the model to accurately predict the pseudo labels, thereby learning to distinguish the different permutations of frequency bands.

Same as spatial jigsaw, learning to solve this, the model captures the critical interrelationships between EEG frequency bands, which significantly influence emotional expression. This pre-text task forms an another cruical part of used self-supervised learning strategy for emotion recognition from EEG data.

4.2.4 Contrastive learning part

Building upon the spatial and frequency jigsaw puzzle tasks, this work also incorporates a contrastive learning approach to further optimize the feature learning process and facilitate the extraction of inherent representations from the EEG data. This approach aims to maximize the similarity between different augmentations of the same EEG emotion data in a shared feature space.

In this context, using data augmentation operation, denoted as Q($\cdot$), which considers both spatial and frequency transformations of the original EEG emotion data. For each original EEG emotion data point Xi, where i ranges from 1 to N, we obtain M augmented versions of this data point: Xi1, Xi2, ..., XiM = Q(Xi). Therefore, each augmented data point is associated with (M - 1) positive pairs and (N - 1) $\times$ M negative pairs.

To map these augmented EEG emotion data points onto the feature space, a projection head is used, denoted as Hp($\cdot$), similar to the SimCLR approach. This process results in Znm = Hp(F(Xnm)), where F($\cdot$) is the shared feature extractor, which is in this case, the GNN.

Contrastive learning approach ensures that positive pairs (i.e., different augmentations of the same EEG data point) are mapped close together in the feature space, while negative pairs (i.e., augmentations of different EEG data points) are mapped far apart. This process not only enhances the model's ability to learn meaningful representations from the EEG data but also significantly improves its emotion recognition performance.

4.3 Multitask learning

Shared feature extractor is leveraged to process EEG data and extract informative features. My first attempts were to build convolutional neural network as the common feature extractor, the backbone for multitask learnin. On the end, I decided to utilize same approach as in [49]. Therefore second extractor is built upon a 1-dimensional Chebyshev convolution (Conv1D Chebyshev), known for its efficiency in processing sequence data like EEG, and capturing both local and global patterns in the data efficiently.

The feature extractor plays a crucial role in the proposed final framework. It works in tandem with the self-supervised tasks and serves as the common backbone for processing and transforming the raw EEG data into a more informative and compact representation. This representation is subsequently used by the distinct classification heads associated with each self-supervised task.

To be specific, I ended having four distinct classification heads, each dedicated to a specific self-supervised task: a spatial head (Hs), a frequency head (Hf), edit head (He), and a projection head (Hp) for the contrastive learning task. Each of these heads is designed to process the features provided by the shared feature extractor and make predictions regarding their respective tasks. The spatial head predicts the permutations of the spatial jigsaw puzzle, the frequency head predicts the permutations of the frequency jigsaw puzzle, edit Head predicts the pseudo-label of edited channel, and the projection head maps the augmented data to a common feature space for the contrastive learning task.

In addition to these, there is also a main classification head. This head is of paramount importance as it is designed to perform the primary task of our study: emotion recognition from EEG data. After being processed by the feature extractor, the EEG data representations are fed into this classification head, which then outputs the predicted emotional states. The classification head is trained on the labeled EEG data, with the objective of minimizing the difference between its predictions and the actual emotional states.

The shared feature extractor and the various classification heads work in synergy, forming an integrated and efficient framework for emotion recognition from EEG data. The feature extractor captures the complex and informative features from the EEG data, while the classification heads leverage these features to perform their designated tasks. This framework not only enables the effective learning of the intricate relationships in the EEG data but also significantly improves the overall emotion recognition performance.

4.4 Graph neural network

Graph-based neural networks (GNNs) and Chebyshev convolution, these elements form a backbone for the model architecture and are particularly well suited to handling the unique structure and complexity of EEG data.

The way this work utilizes the GNN, is that the convolutions are done, throughout the graph. Name of the convolution specifically is ChebConv. It is a graph convolutional operation that uses Chebyshev polynomials to aggregate information from neighboring nodes in a graph, enabling efficient use of neural net architectures for graph-structured data. Parameter K specify the distance traveled across connections. Best results were observed to be at K=2, this should undergo a further study.

The Chebyshev convolution is a spectral convolution operation based on Chebyshev polynomials, which are a set of orthogonal polynomials. This choice is motivated by several factors. Firstly, Conv1D Chebyshev layers are well suited to sequence data like EEG, as they can effectively capture both local and global temporal patterns. They have also been shown to excel at extracting meaningful features from graph-structured data.

Model employs the Conv1D Chebyshev layers within the shared feature extractor to transform raw EEG data into a more informative and compact representation. This representation is then used by the distinct classification heads associated with each self-supervised task: spatial permutation prediction, frequency permutation prediction, edit pseudo-label prediction and contrastive learning. Tasks are designed to encourage the model to learn useful representations by solving challenging auxiliary tasks. They also serve to regularize the model and prevent overfitting.

GNN and the convolution used is the main knowledge base of the whole final model. As it is used as a backbone. Previous classical convolution neural net, has not achieved such a good performance. Backbone is a part of neural network where the most general info about fundamentals of the problem lays. Because of the nature of this field, many tasks have many things different but also similar. So the idea of backbone is for example to train net for emotion prediction on one dataset and then on second dataset. These tasks surely have many aspects similar and share domain specific information. It is a novel and scalable approach to get better accuracy when building neural nets.

The GNN can capture how the emotional state of the individual affects the brain's connectivity pattern. For example, different emotions might be associated with different patterns of connectivity in the brain, and the GNN can learn to recognize these patterns.

Chapter 5

Experiments and Results

This chapter expands on experiments throughout the thesis. It is divided to baseline model one, baseline model two (CNN-FFNN), multi-task CNN-FFNN, and the final one multi-task GNN. All models deals only with SEED dataset.

As said in previous chapter, many features and their combinations were considered. None of them were as suitable as differential entropy x power in frequency bands, as this feature set is used along other models for emotion recognition from EEG, it is a good comparing anchor. So first models used only raw EEG data, next there are models with different kind of inputs and then the most promising ones are based on differential entropy x power in each frequency band. More about this, in further reading.

5.1 Baseline model

First experiments were done with 5ms inputs, in these, not enough information is present. Building my baseline model, the input was just a raw EEG data, 62 values measured in microvolts. Achieved 69% of accuracy, and after few upgrades the accuracy was nice 92% this to me was suprising, even suspicious. I found out that there was a data leakage. So after fixing this up, the model has a problem to overfit even on one batch. The accuracy was 42% and after some upgrades it goes up to 50% accuracy. Consiering only three classes, the score is relatively horrible. For me this was the reason why I quit such input.

To have the model comparable with others, I decided to classify based on 1 second of recording. Such an interval contain enough information to classify emotion from EEG, at least on used datasets. This can be considered as a proof of information present.

First models I have created, those were mid-baseline achieved 50% - 55% accuracy. As the input, differential entropy of each channel times five power in mentioned frequency bands resulting in the shape (62,5). The score, was as I assumed to be. After trying different architectures of FFNN and their combinations, I archieved 56% accuracy. Further optimizing, addding second branch for better gradient propagation achieved 57%, and adding proper residual connection then Improved the score to 58%. On one of the last experiments I noticed that I am doing softmax two times, because of the criterion „CrossEntropyLoss“ already does softmax. This little bugfix improved the score to 59% on SEED dataset. Trying different activation (leaky_ReLU, eLU)functions didn't improve the score. ReLU is the one.

I have tried many settings of the hyperparameters and architectures. None of them is even equally good as the model with CNN. The model could be rebuild to form an ensemble, but I opted to discard this solution. Instead of that I moved to another architecture, with CNN.

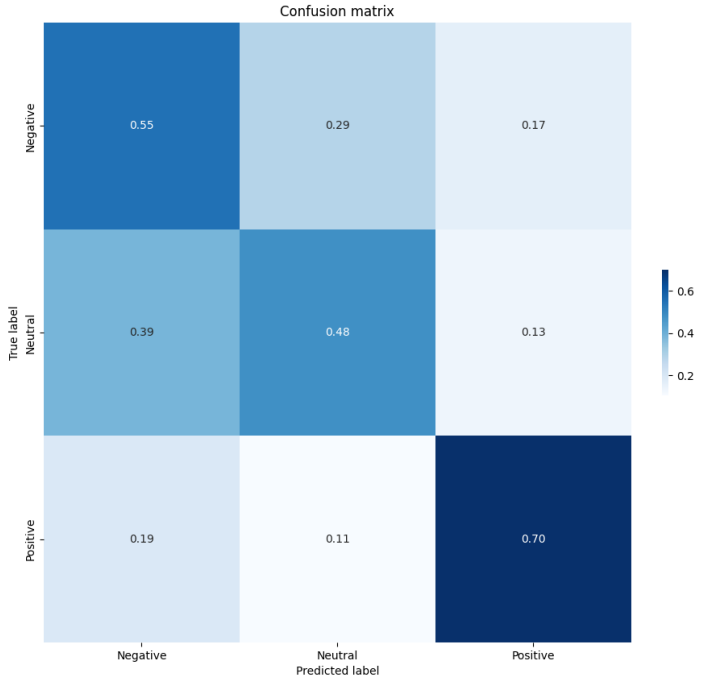

Figure 5.1: Baseline model confusion matrix

The findings derived from the implementation of this model are far from insignificant. An examination of the confusion matrix reveals a distinctly more accurate prediction capability for the category of positive emotions in comparison to the other classifications. This observation, which has repeated over multiple experimental iterations even with balanced dataset, suggests a greater divergence in the model's ability to identify positive emotional states as opposed to other states. This could be attributable to the inherent differences between the mental states themselves. For instance, the disparity between neutral and sad states might be considerably narrower, thereby making them more challenging to distinguish for the model. The positive emotional state appears to be more distinct, yielding a wider gap when compared with both neutral and sad states. This differential characteristic seems to enhance the model's predictive accuracy for positive emotions.

5.2 CNN-FFNN model

The initial architecture of the model was inspired by successful architectures from Kaggle [29] [52]. Utilizing a Convolutional Neural Network (CNN) proved effective in capturing

the required features, and even the first experiment achieved a promising accuracy of 65%. By refining the architecture, an additional increase of 7 percentage points was achieved, bringing the total accuracy to 72%.

Although this score was still not satisfactory, further improvements were made by tuning the hyperparameters, setting up an effective early stopping mechanism, and applying a learning rate scheduler. These modifications helped to improve the model's accuracy to 75%, a figure obtained as an average from three trials.

Dropout values in the range of 0.25 to 0.30 yielded the best results. However, it was crucial to switch the network into evaluation mode (net.eval()) during testing. If not, the training techniques remained active and negatively impacted the results.

While normalizing the data didn't directly enhance the model's accuracy, it accelerated the training phase, contributing to a more efficient experimentation process.

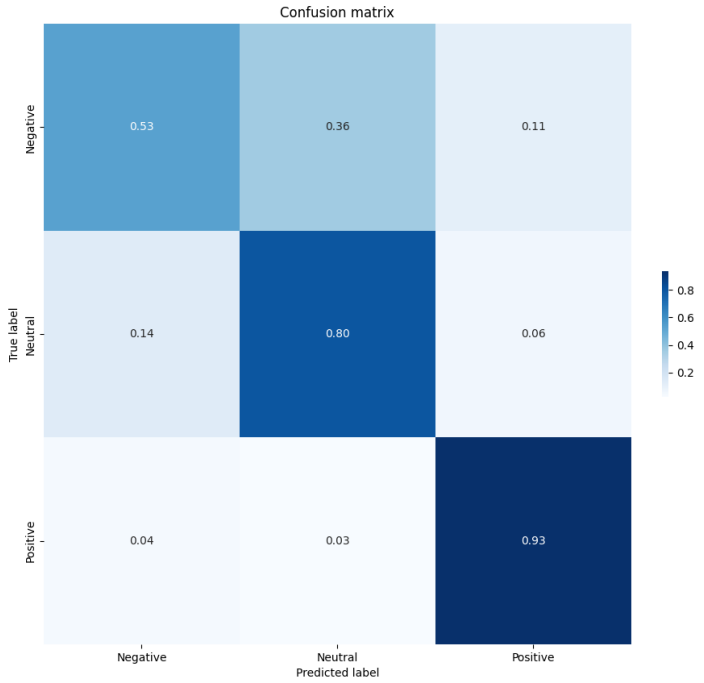

Figure 5.2: CNN-FFNN model confusion matrix

As the complexity of the model increased, it became good at identifying features unique to the neutral emotion. Consequently, the resulting confusion matrix indicates that the model can accurately predict positive and neutral emotions. However, it struggles to distinguish negative emotions. This pattern was consistently observed throughout the experimentation process.

5.3 CNN-FFNN multitask model

This model is the strongest model, build from scratch in this thesis. It utilizes the self supervised learning technique, more than once. The proposed model, has differential entropy cross power in frequency bands as a input. This means that the information about time as well as information from frequency domain is encapsulated into this.

While I was creating self supervised tasks, I was thinking on spatial and connectivity info. So i decided to make tasks tailored to my needs. So tasks I created are random noise injection and random shuffling in both dimensions.

Noise injection The noise injection task is built on the premise that predicting a slightly distorted channel should provide insights beneficial to emotion prediction. To further exploit this potential, the task was incorporated into the model, functioning in a multi-task manner. This approach enriches the common feature extractor by incorporating more diverse learning signals.

Channel shuffling To further capture the connectivity and spatial information inherent in the data, another self-supervised task was introduced, focused on permutation prediction. In this task, the order of the 62 channels is randomly shuffled. Each specific permutation is assigned a unique label, and a separate classification head is trained to predict this label, thereby learning the original order of the channels.

The number of possible permutations for 62 channels (62 factorial) is astronomically large, making it impractical for direct permutation prediction. To circumvent this issue, a subset of permutations was selected for the task.

Frequency shuffling To further leverage the data, now from another angle, a similar task to previous channel shuffling was introduced in the frequency domain. Given that there are only five frequency bands, the total number of permutations is 5 factorial, a number manageable enough for direct use by the classifier.

5.3.1 Noise injection Results

This task proved to be highly effective, capturing vital information and introducing regularization into the model. This regularization made it more difficult for the model to overfit to the original data and task. Adjustments were made to early stopping and the scheduler to better suit the model's needs. Increasing patience for early stopping did not yield significant improvements. However, adjustments to the scheduler resulted in more pronounced positive outcomes.

One challenge encountered during development was that the gradient from the self-supervised task significantly exceeded the gradient from the emotion classification head. To address this issue, the gradients had to be adjusted throughout the training.

The model is trained from shuffled data of all recordings, making first 9 trials of recording for each person as training data, and the rest (6 recordings) as a testing set. Yielded not

so good training process. The self supervised task effectively tackled this problem as can be seen in the picture 5.5. Still a high overfit to current batch is present. Confusion matrix after adding noise injection self supervised task is in the picture 5.3, achieved accuracy on test data is 77%.

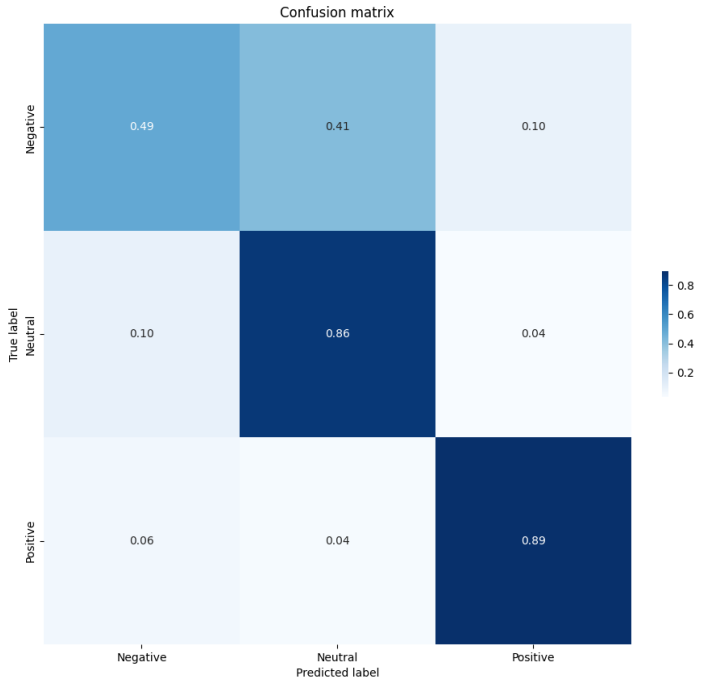

Figure 5.3: Confusion matrix after noise injection self supervised task

Figure 5.4: Before noise injection self supervised task

Figure 5.5: After noise injection self supervised task

5.3.2 Channel shuffling Results

This approach led to a satisfactory learning curve for the permutation prediction task and complemented the learning curve of the main emotion classification task effectively. Meaning the peaks, from previous, have vanished. Overfitting has decreased 5.6. Please note that the number of epoch and the loss score was varied due to different setting of a loss function and the scheduler parameters.

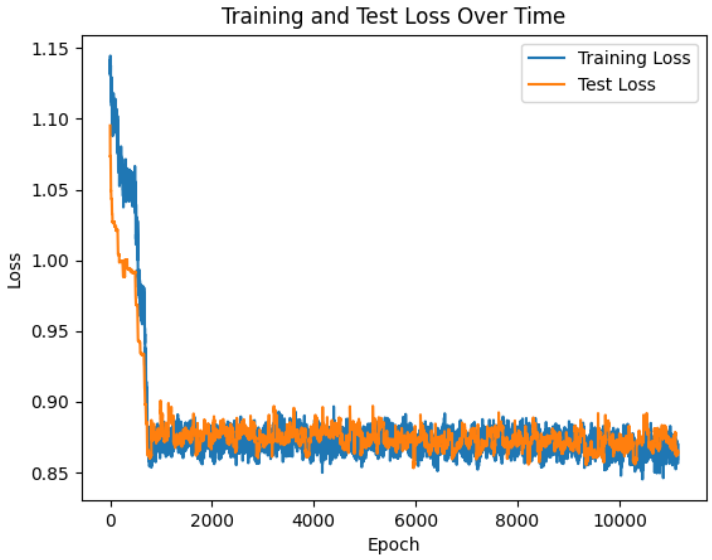

Figure 5.6: After channel shuffle self supervised task training process

However, it should be noted that the inclusion of these modifications did not result in noticeable improvements to the final test accuracy or the balance of the confusion matrix. Further investigation is necessary to determine the optimal approach to incorporate these methods and gain the most advantage from them, like in the [49] where rather that shuffling 62 channel, it does only a shuffling of 10 brain regions.

While the implementation of this pretext task has indeed led to some improvements, it's important to note that it hasn't made significant strides in addressing the challenge of low accuracy scores for negative emotions. Despite my hopes and efforts, this approach has not provided the anticipated breakthrough in this specific area.

5.3.3 Frequency shuffling Results

This self-supervised task represents a significant milestone, primarily because it has proven effective in tackling the previously observed issue of low accuracy in predicting negative emotions 5.7. For the first time, the model was partly successful (across many trials) with this task. By extending the concept of permutation prediction to the frequency domain, it has enabled the model to gain a deeper comprehension of the patterns within and across frequency bands. This enhanced understanding has, in turn, led to an overall improvement in the model's performance by 1%.

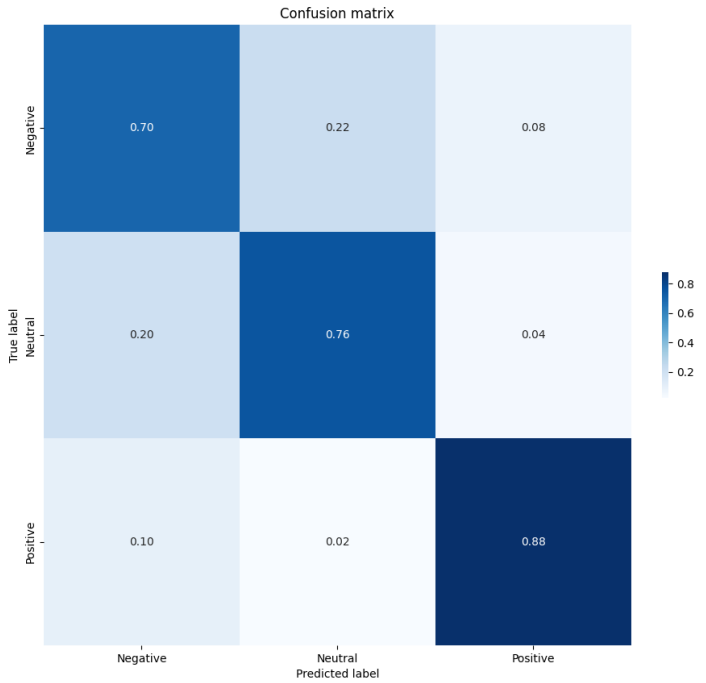

Figure 5.7: Confusion matrix after frequency shuffle self supervised task

5.3.4 Further improvements

The model was developed iteratively, and observing its tendency to overfit on a single batch led me to explore methods of improvement. This process resulted in a set of beneficial parameters which will be briefly recapitulated. Note that all training was performed using CUDA.

Normalization using classical mean/standard deviation proved most effective, primarily in terms of learning speed. Varying batch sizes occasionally led to more holistic model behavior. This approach was quite experimental, but proved beneficial in some instances. The settings for the scheduler were slightly indulgent, with ReduceLROnPlateau set to a patience of 17 and EarlyStopping's patience set to 30.

Both undersampling and oversampling were considered; however, they did not yield significant improvements. Unfortunatelly it does not help with the low accuracy on negative emotions either, even with all classes balanced. Another issue was the model's slow start. To address this, I employed weight initialization methods such as Xavier and Kaiming. While these methods did not contribute to the overall accuracy score on the test set, they did help mitigate the slow start, as observed across multiple trials.

Various optimizers such as RMSprop, Momentum, and Adam were experimented with, with Adam ultimately selected due to its stable performance throughout the experimentation. The learning rate was set to e to power of -3 and weight decay to e to power of -5. Each time the model achieved its best results on the test set, the weights of the best model

were saved and loaded for use in the next epoch. This approach allowed for exploration in each epoch while saving the best model provided a basis for exploitation.

I also experimented with varying the loss function and dropout settings throughout the training process. For example, after certain batches, the dropout was changed from 0.25 to 0.8, among other values. Similarly, the loss function, which generally served as a composite loss of four losses (the original emotion prediction loss, and losses from the three other self-supervised tasks), was adjusted by changing the priority of different tasks throughout training. This resulted in an improved accuracy score, indicating its potential benefit for further study.

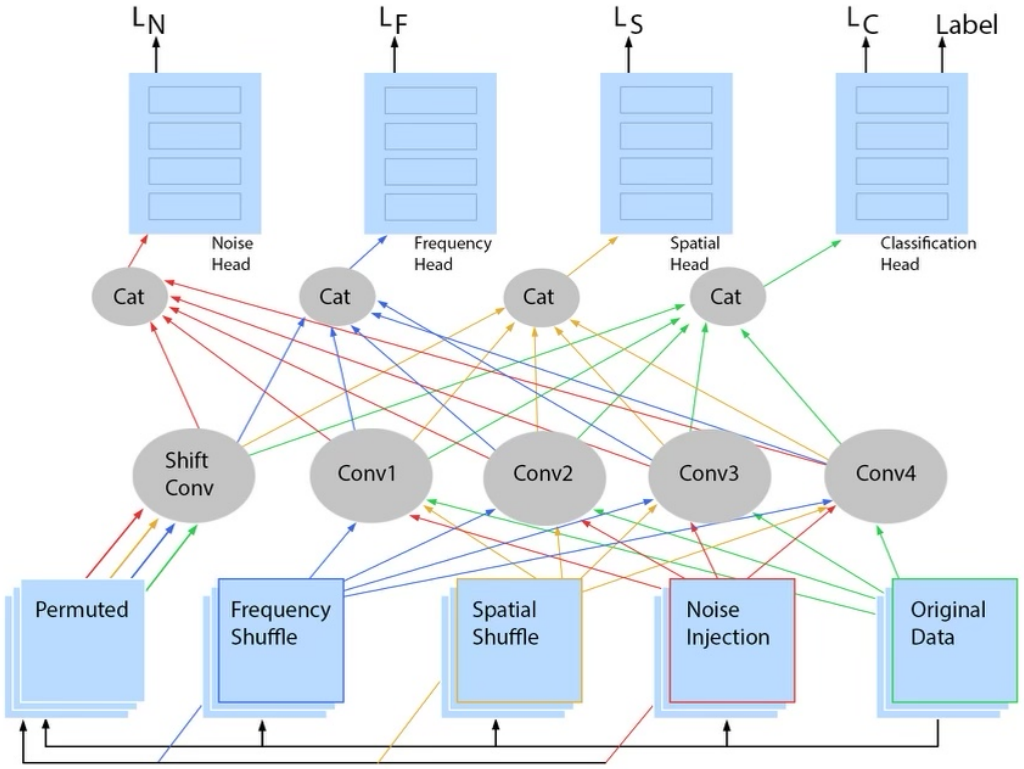

Figure 5.8: CNN multitask model architecture

Architecture Finally, I would like to describe the evolution of my model's architecture. It began as a baseline convolutional neural network (CNN) model, complemented by a linear part. In the first upgrade, a noise classification head, identical in architecture to the main classification head, was added.

The next step was to incorporate two additional self-supervised tasks, following the same approach as with the noise head. This quadrupled the input data, with one set for each self-supervised head. The backbone or common feature extractor of the model is the CNN.

In the final iteration, more common feature extractors were added to capture different types of information (global, local) via varying kernel sizes, padding, and strides. Despite being more computationally expensive, this approach was effective in extracting more complex information from the data. The final model achieved an accuracy of 80% on the test data, with a total F1 score of 0.77 and a total micro AUC of 0.91. The confusion matrix for the final model can be seen in Figure 5.11.

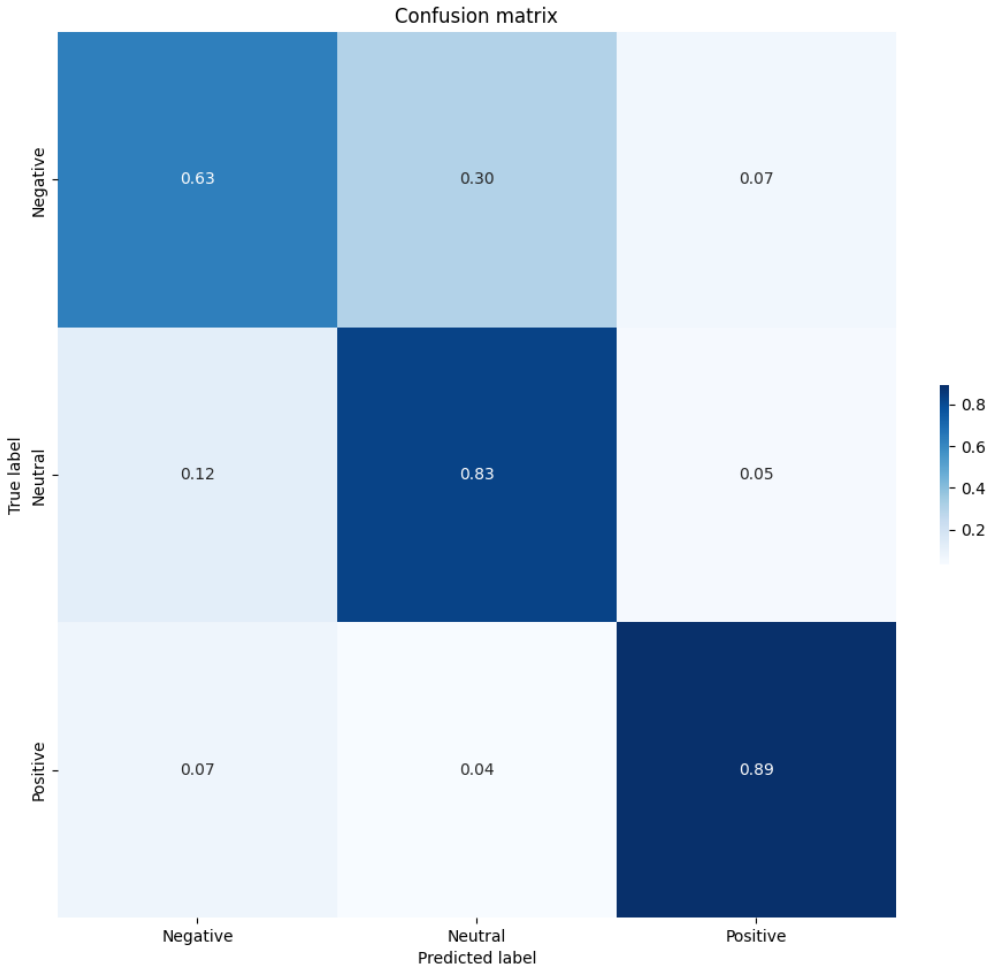

Figure 5.9: Confusion matrix of my model after all changes

5.4 GMSS

The referenced model [49], which utilizes a Graph Neural Network (GNN) rather than a traditional Convolutional Neural Network (CNN), has achieved state-of-the-art results on the SEED and SEED IV datasets. A key aspect of this model's success lies in its use of contrastive learning and a projection head. These techniques serve to move similar sam-

ples closer in feature space while distancing dissimilar ones, contributing to its impressive accuracy of 96.48%.

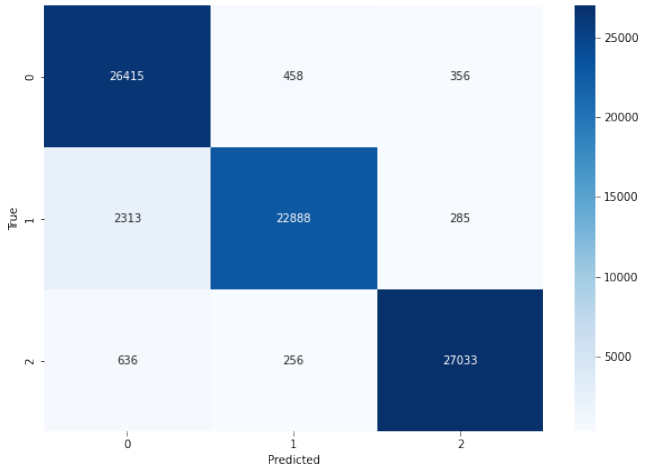

Figure 5.10: GMSS original model confusion matrix

Despite the difference in the underlying neural network, this model shares similarities with my own, particularly in the incorporation of self-supervised shuffling tasks. Its approach to the spatial task differs slightly, dividing the channels into ten groups before shuffling. However, the total number of permutations, 10 factorial, is still prohibitively large for direct prediction. Hence, similar to my own approach, this model selects a subset of permutations for the task.

To enhance the performance of GMSS [49] model, I initiated several attempts at varying the architecture and hyperparameters. The first modification involved implementing normalization to the model's input data. While this did not lead to a direct improvement in the model's performance, it did expedite the training process to some degree, making it more time-efficient.

Simultaneously, I experimented with various hyperparameters such as the learning rate, batch size, and number of epochs, but these modifications did not yield noticeable improvements in the model's performance.

Further, modifications were explored by varying the model's activation function. I experimented with different types of activation functions such as the Leaky Rectified Linear Unit (Leaky ReLU) and the Exponential Linear Unit (ELU). However, these changes also did not lead to significant performance enhancements.

In addition to these, I tried altering the model's architecture by varying the number of layers and neurons per layer, as well as experimenting with different types of layers such as convolutional, pooling, and fully-connected layers. I also attempted implementing different types of regularization techniques such as L1 and L2 regularization, dropout, and batch normalization. Furthermore, I experimented with different optimization algorithms beyond stochastic gradient descent, such as Adam, RMSprop, and Adagrad or momentum.

Despite these exhaustive attempts, the model's performance remained largely unchanged. This is likely due to the inherent complexity of the task and the high initial performance of the model. The impressive starting score suggests that the model has already captured a significant amount of the available information, making further improvements challenging. Any alterations, therefore, must be made carefully, as they risk disrupting the finely tuned balance the model has achieved.

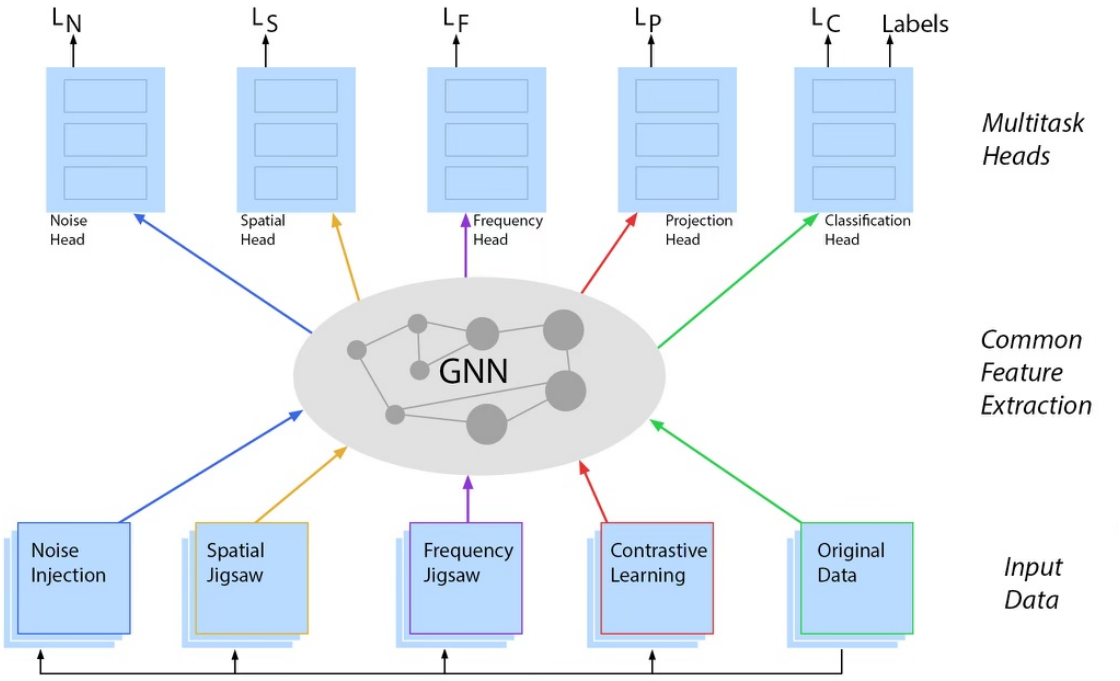

Figure 5.11: GMSS model architecture

My main contribution is that I managed to further improve this results. The model has already showed a very promising results, thanks to many things. Graph neural network to capture graph structured insights, informative feature set, self supervised tasks for exploitation of spatial and frequency domain, contrastive learning. But one of the things that it lacks is the anomaly detection, to tackle this I decided to integrate my self supervised task here. The task where one channel is edited and the model tries to predict which one of the channel has been edited. So after this the model yielded 99.7% accuracy on testing data (average after 5 trials). The score is that high, that it is even questionable, but on the other hand the overall model architecture and details are strong mechanism with capacity to do this. Resulting in the state of the art result on SEED dataset following the same experimental protocol for dependent testing as the compared solutions. The confusion matrix shows the success of adding my self supervised task.

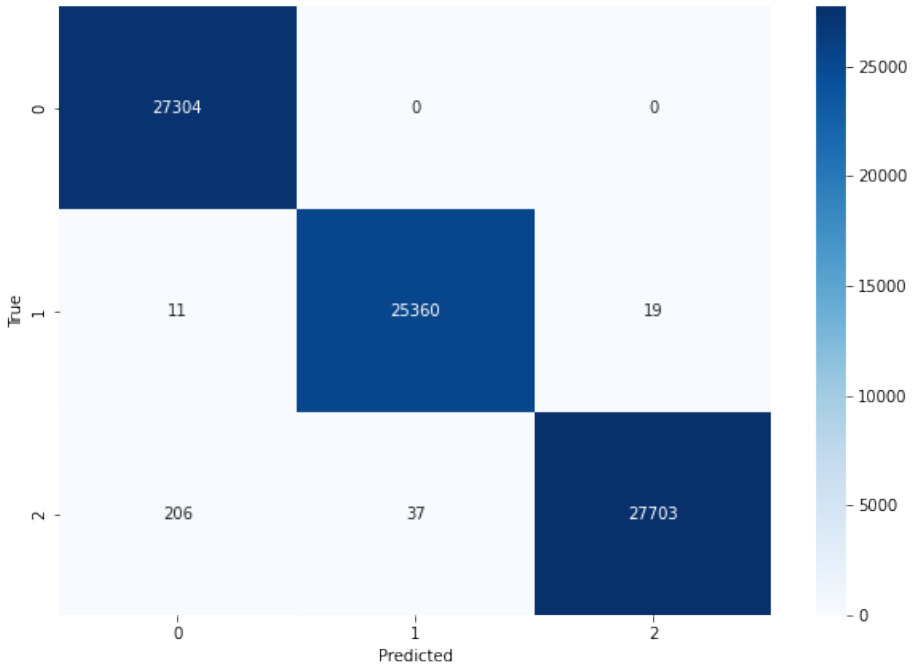

Figure 5.12: GMSS upgraded model confusion matrix

Chapter 6

Conclusion

In this thesis, I embarked on the challenging yet rewarding journey of leveraging self-supervised learning strategies to enhance emotion recognition from EEG data. The primary aim was to develop a model with improved discriminatory abilities by harnessing the spatial and frequency patterns inherent in the EEG signals. This work holds substantial significance, given the immense potential applications of emotion recognition in fields ranging from healthcare to human-computer interaction.

My initial investigation led me to a Convolutional Neural Network (CNN) model. Iterative refinement of this model provided valuable insights into the nature of the data and the design considerations for such models. Notably, the CNN proved capable of classifying emotions based on differential entropy and power in frequency bands. The architecture of the model emerged as a critical factor, with a particular emphasis on good gradient passing and meaningful feature extraction.

Recognizing that EEG recordings inherently present data that can be understood as a graph, I explored the use of a Graph Neural Network (GNN). This model proved superior in harnessing information effectively, as it naturally aligns with the spatial interdependencies present in EEG data.

The design and implementation of self-supervised tasks were crucial aspects of this research. My initial approach to a spatial self-supervised task involved shuffling all channels. However, a more refined approach, as demonstrated in the GMSS [49] study, grouped the channels into 10 categories corresponding to brain regions. This approach proved to be more aligned with the primary task of emotion classification.

Furthermore, the introduction of an anomaly detection self-supervised task presented a significant advancement in our model. The implementation of this task on the already robust GMSS [49] model (SOTA) led to a remarkable improvement in performance, with the accuracy escalating from 96.4% to 99.7%. This result underscores the immense potential of self-supervised learning tasks in enhancing model performance.

Modifying the settings during the training phase, specifically adjusting the dropout rate and loss function, resulted in a improvement of approximately 2-3% in accuracy. This enhancement underscores the importance of tuning and optimization in the model training process. This upgrade shows to be stable across many trials.

The integration of contrastive learning emerged as another significant enhancement to the model. This self-supervised task, which focuses on learning by comparison, helped the model to learn more discriminative features. However, the success of contrastive learning is strongly dependent on the selection of positive and negative examples, which directly impact the quality of the learned representations.

Despite the promising results, this study is not without limitations. The choice of architecture, self-supervised tasks, and training settings can always be further refined. Moreover, the inherently noisy nature of EEG data presents ongoing challenges for model performance.

Looking forward, there are several potential avenues for further research. These include exploring other architectures and self-supervised tasks, devising more effective strategies for example selection in contrastive learning, and investigating other representations of EEG data. This study has laid a solid foundation for such future endeavors, paving the way for more refined models for emotion recognition from EEG data.

Bibliography

[1] *Action potential image.* Available at: https://www.khanacademy.org/science/biology/human-biology/neuron-nervous-system/a/the-synapse.

[2] *Brain lobes and function image.* Available at: https://www.grepmed.com/images/5344/cerebral-function-localization-lobes-deficits.

[3] *Brain stem image.* Available at: https://commons.wikimedia.org/wiki/File:1311_Brain_Stem.jpg.

[4] *Classical Machine Learning pipeline.* Available at: https://www.altexsoft.com/blog/machine-learning-pipeline/.

[5] *Confusion Matrix.* Available at: https://towardsdatascience.com/visual-guide-to-the-confusion-matrix-bb63730c8eba.

[6] *Corpus callosum.* Available at: http://cosmology.com/Consciousness163.html.

[7] *Corpus callosum second picture.* Available at: https://www.researchgate.net/figure/The-general-organization-of-the-corpus-callosum-A-C-The-structure-and-location-of-the_fig1_344412345.

[8] *Differences between ECoG, EEG and Implants.* Available at: https://www.bmseed.com/stretchable-electrodes-for-brain-implants.

[9] *Emotion faces.* Available at: https://www.anxiety.org/failure-to-process-emotional-faces-can-cause-anxiety-and-social-disorders.

[10] *Examples of visual stimuli image.* Available at: https://www.researchgate.net/figure/Examples-from-the-IAPS-to-evoke-specific-emotions-48_fig3_333061423.

[11] *K-nearest neighbours image.* Available at: https://www.ibm.com/topics/knn.

[12] *Kaggle competition won using cnn ensemble.* Available at: https://www.kaggle.com/competitions/grasp-and-lift-eeg-detection/leaderboard.

[13] *Lobes.* Available at: https://www.mdpi.com/2079-9292/10/23/3037/htm#B23-electronics-10-03037.

[14] *LSTM and GRU informations.* Available at: https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21.

[15] *LSTM GRU RNN compared.* Available at: http://dprogrammer.org/rnn-lstm-gru.

[16] *Montage setting picture*. Available at: https://www.learningeeg.com/montages-and-technical-components.

[17] *Neuron cell image*. Available at: https://my-ms.org/anatomy_microanatomy.htm.

[18] *ROC curve*. Available at: https://en.wikipedia.org/wiki/Receiver_operating_characteristic#/media/File:Roc_curve.svg.

[19] *Spatial and temporal differencies between different modalities*. Available at: https://www.researchgate.net/post/How-to-decide-whether-to-use-EEG-MEG-or-fMRI.

[20] *Spectral edge frequency picture*. Available at: https://www.researchgate.net/publication/312462395_Automatic_Sleep_Monitoring_Using_Ear-EEG.

[21] *Valence Arousal Dominance emotion continuous model*. Available at: https://www.researchgate.net/figure/The-VAD-Valence-Arousal-Dominance-model-spanned-across-the-six-basic-emotions_fig1_338118399.

[22] *Valence arousal measure*. Available at: https://www.researchgate.net/figure/Two-dimensional-valence-arousal-space_fig1_304124018.

[23] *Functional near-infrared spectroscopy*. Wikimedia Foundation, Dec 2022. Available at: https://en.wikipedia.org/wiki/Functional_near-infrared_spectroscopy.

[24] *Corpus callosum*. Wikimedia Foundation, Jan 2023. Available at: https://en.wikipedia.org/wiki/Corpus_callosum.

[25] Ahangi, A. *Using Combining Classifiers in Brain-Computer Interfacing*. Dissertation.

[26] Averill, J. R. Individual Differences in Emotional Creativity: Structure and Correlates. *Journal of Personality*. Wiley. february 1999, vol. 67, no. 2, p. 331–371. DOI: 10.1111/1467-6494.00058. Available at: https://doi.org/10.1111/1467-6494.00058.

[27] Ayata, D., Yaslan, Y. and Kamaşak, M. Emotion recognition via random forest and galvanic skin response: Comparison of time based feature sets, window sizes and wavelet approaches. In: *2016 Medical Technologies National Congress (TIPTEKNO)*. 2016, p. 1–4. DOI: 10.1109/TIPTEKNO.2016.7863130.

[28] Bahari, F. and Janghorbani, A. EEG-based emotion recognition using Recurrence Plot analysis and K nearest neighbor classifier. In: *2013 20th Iranian Conference on Biomedical Engineering (ICBME)*. 2013, p. 228–233. DOI: 10.1109/ICBME.2013.6782224.

[29] Banggiangle, B. G. L. *CNN EEG PYTORCH*. Kaggle, Apr 2021. Available at: https://www.kaggle.com/code/banggiangle/cnn-eeg-pytorch.

[30] Biazoli, C. E., Sturzbecher, M., White, T. P., Santos Onias, H. H. dos, Andrade, K. C. et al. Application of Partial Directed Coherence to the Analysis of Resting-State EEG-fMRI Data. *Brain Connectivity*. Mary Ann Liebert Inc. december 2013, vol. 3, no. 6, p. 563–568. DOI: 10.1089/brain.2012.0135. Available at: https://doi.org/10.1089/brain.2012.0135.

[31] Bradley, M. M. and Lang, P. J. International Affective Picture System. In: *Encyclopedia of Personality and Individual Differences*. Springer International Publishing, 2017, p. 1–4. Available at: https://doi.org/10.1007/978-3-319-28099-8_42-1.

[32] Clercq, W., Vergult, A., Vanrumste, B., Paesschen, W. and Huffel, S. Canonical Correlation Analysis Applied to Remove Muscle Artifacts From the Electroencephalogram. *IEEE transactions on bio-medical engineering*. january 2007, vol. 53, p. 2583–7. DOI: 10.1109/TBME.2006.879459.

[33] Darwin, C., Ekman, P. and Prodger, P. *The Expression of the Emotions in Man and Animals*. Oxford University Press, 1998. ISBN 9780195158069. Available at: https://books.google.cz/books?id=TFRtLZSHMcYC.

[34] De Clercq, W., Vergult, A., Vanrumste, B., Van Paesschen, W. and Van Huffel, S. Canonical Correlation Analysis Applied to Remove Muscle Artifacts From the Electroencephalogram. *IEEE Transactions on Biomedical Engineering*. 2006, vol. 53, no. 12, p. 2583–2587. DOI: 10.1109/TBME.2006.879459.

[35] Ekman, P. Are there basic emotions? *Psychological Review*. American Psychological Association (APA). 1992, vol. 99, no. 3, p. 550–553. DOI: 10.1037/0033-295x.99.3.550. Available at: https://doi.org/10.1037/0033-295x.99.3.550.

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

[37] Elbert, T., Lutzenberger, W., Rockstroh, B., Berg, P. and Cohen, R. Physical aspects of the EEG in schizophrenics. *Biological Psychiatry*. 1992, vol. 32, no. 7, p. 595–606. DOI: https://doi.org/10.1016/0006-3223(92)90072-8. ISSN 0006-3223. Available at: https://www.sciencedirect.com/science/article/pii/0006322392900728.

[38] Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*. Springer Science and Business Media LLC. april 1980, vol. 36, no. 4, p. 193–202. DOI: 10.1007/bf00344251. Available at: https://doi.org/10.1007/bf00344251.

[39] Guo, J. Backpropagation through time. *Unpubl. ms., Harbin Institute of Technology*. 2013.

[40] He, K., Zhang, X., Ren, S. and Sun, J. *Deep Residual Learning for Image Recognition*. arXiv, 2015. DOI: 10.48550/ARXIV.1512.03385. Available at: https://arxiv.org/abs/1512.03385.

[41] Hjorth, B. EEG analysis based on time domain properties. *Electroencephalography and Clinical Neurophysiology*. 1970, vol. 29, no. 3, p. 306–310. DOI: https://doi.org/10.1016/0013-4694(70)90143-4. ISSN 0013-4694. Available at: https://www.sciencedirect.com/science/article/pii/0013469470901434.

[42] Ho, T. K. Random decision forests. In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. 1995, vol. 1, p. 278–282 vol.1. DOI: 10.1109/ICDAR.1995.598994.

[43] Houssein, E. H., Hammad, A. and Ali, A. A. Human emotion recognition from EEG-based brain–computer interface using machine learning: a comprehensive review. *Neural Computing and Applications*. Springer Science and Business Media LLC. may 2022, vol. 34, no. 15, p. 12527–12557. DOI: 10.1007/s00521-022-07292-4. Available at: https://doi.org/10.1007/s00521-022-07292-4.

[44] Huang, D., Ren, A., Shang, J., Lei, Q., Zhang, Y. et al. Combining Partial Directed Coherence and Graph Theory to Analyse Effective Brain Networks of Different Mental Tasks. *Frontiers in Human Neuroscience*. Frontiers Media SA. may 2016, vol. 10. DOI: 10.3389/fnhum.2016.00235. Available at: https://doi.org/10.3389/fnhum.2016.00235.

[45] Jiang, X., Bian, G.-B. and Tian, Z. Removal of artifacts from EEG Signals: A Review. *Sensors*. 2019, vol. 19, no. 5, p. 987. DOI: 10.3390/s19050987.

[46] Kendall, A., Gal, Y. and Cipolla, R. *Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics*. 2018.

[47] Kleinow, T. *Testing continuous time models in financial markets*. 2002. Dissertation. Humboldt-Universität zu Berlin, Wirtschaftswissenschaftliche Fakultät.

[48] Kolen, J. F. and Kremer, S. C. Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies. In: *A Field Guide to Dynamical Recurrent Networks*. 2001, p. 237–243. DOI: 10.1109/9780470544037.ch14.

[49] Li, Y., Chen, J., Li, F., Fu, B., Wu, H. et al. *GMSS: Graph-Based Multi-Task Self-Supervised Learning for EEG Emotion Recognition*. 2022.

[50] Li, Y., Wang, L., Zheng, W., Zong, Y., Qi, L. et al. A Novel Bi-Hemispheric Discrepancy Model for EEG Emotion Recognition. *IEEE Transactions on Cognitive and Developmental Systems*. 2021, vol. 13, no. 2, p. 354–367. DOI: 10.1109/TCDS.2020.2999337.

[51] Li, Y., Zheng, W., Wang, L., Zong, Y., Qi, L. et al. *A Novel Bi-hemispheric Discrepancy Model for EEG Emotion Recognition*. 2019.

[52] m0nika, M. S. *Seizurecnn*. Kaggle, May 2020. Available at: https://www.kaggle.com/code/m0nika/seizurecnn.

[53] Mauss, I. B. and Robinson, M. D. Measures of emotion: A review. *Cognition &amp Emotion*. Informa UK Limited. february 2009, vol. 23, no. 2, p. 209–237. DOI: 10.1080/02699930802204677. Available at: https://doi.org/10.1080/02699930802204677.

[54] Mehmood, R. M. and Lee, H. J. Emotion classification of EEG brain signal using SVM and KNN. In: *2015 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*. 2015, p. 1–5. DOI: 10.1109/ICMEW.2015.7169786.

[55] Menick, J., Elsen, E., Evci, U., Osindero, S., Simonyan, K. et al. *A Practical Sparse Approximation for Real Time Recurrent Learning*. arXiv, 2020. DOI: 10.48550/ARXIV.2006.07232. Available at: https://arxiv.org/abs/2006.07232.

[56] Naji, M., Firoozabadi, M. and Azadfallah, P. Emotion classification during music listening from forehead biosignals. *Signal, Image and Video Processing*. Springer Science and Business Media LLC. december 2013, vol. 9, no. 6, p. 1365–1375. DOI: 10.1007/s11760-013-0591-6. Available at: https://doi.org/10.1007/s11760-013-0591-6.

[57] Nolte, G., Ziehe, A., Nikulin, V., Schlögl, A., Krämer, N. et al. Robustly Estimating the Flow Direction of Information in Complex Physical Systems. *Physical Review Letters*. June 2008, vol. 100, p. 234101. Available at: http://doc.ml.tu-berlin.de/causality.

[58] Peterson, L. E. K-nearest neighbor. *Scholarpedia*. 2009, vol. 4, no. 2, p. 1883. DOI: 10.4249/scholarpedia.1883. revision #137311.

[59] Pincus, S. Approximate entropy (ApEn) as a complexity measure. *Chaos (Woodbury, N.Y.)*. april 1995, vol. 5, p. 110–117. DOI: 10.1063/1.166092.

[60] Pinker, S. *How the Mind Works*. Norton, 1997. ISBN 9780393045352. Available at: https://books.google.cz/books?id=fBaqQgAACAAJ.

[61] Reddy, B. K. and Delen, D. Predicting hospital readmission for lupus patients: An RNN-LSTM-based deep-learning methodology. *Computers in Biology and Medicine*. Elsevier BV. october 2018, vol. 101, p. 199–209. DOI: 10.1016/j.compbiomed.2018.08.029. Available at: https://doi.org/10.1016/j.compbiomed.2018.08.029.

[62] Russell, J. A. A circumplex model of affect. *Journal of Personality and Social Psychology*. American Psychological Association (APA). december 1980, vol. 39, no. 6, p. 1161–1178. DOI: 10.1037/h0077714. Available at: https://doi.org/10.1037/h0077714.

[63] Russell, J. A. A circumplex model of affect. *Journal of Personality and Social Psychology*. American Psychological Association (APA). december 1980, vol. 39, no. 6, p. 1161–1178. DOI: 10.1037/h0077714. Available at: https://doi.org/10.1037/h0077714.

[64] Sabeti, M., Katebi, S. and Boostani, R. Entropy and complexity measures for EEG signal classification of schizophrenic and control participants. *Artificial Intelligence in Medicine*. 2009, vol. 47, no. 3, p. 263–274. DOI: https://doi.org/10.1016/j.artmed.2009.03.003. ISSN 0933-3657. Available at: https://www.sciencedirect.com/science/article/pii/S0933365709000530.

[65] Sheng, H. and Chen, Y. *Multifractional Property Analysis of Human Sleep EEG Signals*. August 2011, Volume 3: 2011 ASME/IEEE International Conference on Mechatronic and Embedded Systems and Applications, Parts A and B, p. 323–328. International Design Engineering Technical Conferences and Computers and Information in Engineering Conference.

[66] Song, T., Zheng, W., Song, P. and Cui, Z. EEG Emotion Recognition Using Dynamical Graph Convolutional Neural Networks. *IEEE Transactions on Affective Computing*. 2020, vol. 11, no. 3, p. 532–541. DOI: 10.1109/TAFFC.2018.2817622.

[67] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*. 2014, vol. 15, no. 56, p. 1929–1958. Available at: http://jmlr.org/papers/v15/srivastava14a.html.

[68] Tomkins, S. S. and Karon, B. P. *Affect, imagery, consciousness / by Silvan S. Tomkins ; with the editorial assistance of Bertram P. Karon*. Springer Pub. Co New York, 1962. 4 v. : p. ISBN 0826105440 0826105432.

[69] Volf, P., Stehlik, M., Kutilek, P., Kloudova, G., Rusnakova, K. et al. Brain Electrical Activity Mapping in Military Pilots During Simulator Trainings. In: *2019 International Conference on Military Technologies (ICMT)*. 2019, p. 1–6. DOI: 10.1109/MILTECHS.2019.8870112.

[70] WHISSELL, C. M. Chapter 5 - THE DICTIONARY OF AFFECT IN LANGUAGE. In: Plutchik, R. and Kellerman, H., ed. *The Measurement of Emotions*. Academic Press, 1989, p. 113–131. DOI: https://doi.org/10.1016/B978-0-12-558704-4.50011-6. ISBN 978-0-12-558704-4. Available at: https://www.sciencedirect.com/science/article/pii/B9780125587044500116.

[71] Yang, W., Makita, K., Nakao, T., Kanayama, N., Machizawa, M. G. et al. Affective auditory stimulus database: An expanded version of the International Affective Digitized Sounds (IADS-E). *Behavior Research Methods*. Springer Science and Business Media LLC. march 2018, vol. 50, no. 4, p. 1415–1429. DOI: 10.3758/s13428-018-1027-6. Available at: https://doi.org/10.3758/s13428-018-1027-6.

[72] Yang, W., Makita, K., Nakao, T., Kanayama, N., Machizawa, M. G. et al. Affective auditory stimulus database: An expanded version of the International Affective Digitized Sounds (IADS-E). *Behavior Research Methods*. Springer Science and Business Media LLC. march 2018, vol. 50, no. 4, p. 1415–1429. DOI: 10.3758/s13428-018-1027-6. Available at: https://doi.org/10.3758/s13428-018-1027-6.

[73] Zheng, W.-L. and Lu, B.-L. Investigating Critical Frequency Bands and Channels for EEG-based Emotion Recognition with Deep Neural Networks. *IEEE Transactions on Autonomous Mental Development*. IEEE. 2015, vol. 7, no. 3, p. 162–175. DOI: 10.1109/TAMD.2015.2431497.

[74] Zhong, P., Wang, D. and Miao, C. *EEG-Based Emotion Recognition Using Regularized Graph Neural Networks*. 2020.

Appendix A

Uploaded files

This chapter describes the contents of my program

Baseline FFNN CNN My baseline FFNN CNN model 75% accuracy

CNN multitask model My model 80% accuracy

GMSS SOTA upgraded model 99.7% accuracy

First atempts Contains some first models created

Time analysis Time analysis codes

Frequency analysis Frequency analysis codes

Source analysis Source analysis codes

Connectivity analysis Connectivity analysis codes

Microstate analysis Microstate analysis codes

STFT, Wavelets, Curvelets Time frequency analysis codes