UNIVERSITY OF ZAGREB
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

MASTER THESIS No. 408

EXPLAINABILITY OF MACHINE LEARNING MODELS IN PREDICTION OF AFFECTIVE DISORDERS

Fani Sentinella-Jerbić

Zagreb, June 2024

UNIVERSITY OF ZAGREB
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

MASTER THESIS No. 408

EXPLAINABILITY OF MACHINE LEARNING MODELS IN PREDICTION OF AFFECTIVE DISORDERS

Fani Sentinella-Jerbić

Zagreb, June 2024

UNIVERSITY OF ZAGREB
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Zagreb, 04 March 2024

MASTER THESIS ASSIGNMENT No. 408

Student:	**Fani Sentinella-Jerbić (0036514645)**
Study:	Computing
Profile:	Data Science
Mentor:	prof. Mario Cifrek

Title: **Explainability of machine learning models in prediction of affective disorders**

Description:

Depression belongs to affective disorders, which are characterized by a drop in vital energy, loss of will and a sad mood. The mechanisms of the occurrence of affective disorders are still unknown, and researching them makes it possible to achieve better patient care. As part of the work, it is necessary to study literature related to machine learning models used in the classification and analysis of EEG signals, and investigate in detail the algorithms used for the explainability of machine learning model predictions. Study the available dataset of EEG signals with labeled diagnoses of affective disorders and extracted features. Make a statistical analysis of the available extracted features. Carry out the classification of affective disorders using machine learning models. Perform feature selection and test the obtained classification models on unseen data from the available dataset. Implement selected algorithms for the explainability of machine learning model predictions and analyze the results. For detailed information, contact Eda Jovičić, mag. ing.

Submission date: 28 June 2024

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 4. ožujka 2024.

DIPLOMSKI ZADATAK br. 408

Pristupnica:	**Fani Sentinella-Jerbić (0036514645)**
Studij:	Računarstvo
Profil:	Znanost o podacima
Mentor:	prof. dr. sc. Mario Cifrek

Zadatak: **Objašnjivost modela strojnog učenja u predikciji afektivnih poremećaja**

Opis zadatka:

Depresija spada u afektivne poremećaje raspoloženja. Karakteriziraju je pad životne energije, gubitak volje i tužno raspoloženje. Mehanizmi nastanka afektivnih poremećaja još uvijek su nepoznati, a njihovim istraživanjem moguće je ostvariti bolju skrb pacijentima. U sklopu rada potrebno je proučiti literaturu vezanu za modele strojnog učenja korištenih u klasifikaciji i analizi signala EEG-a, te detaljno istražiti algoritme korištene za objašnjivost predikcija modela strojnog učenja. Proučiti dostupnu bazu podataka signala EEG-a s označenim dijagnozama afektivnih poremećaja i izdvojenim značajkama. Napraviti statističku analizu dostupnog skupa značajki. Provesti klasifikaciju afektivnih poremećaja modelima strojnog učenja. Provesti odabir značajki, te testirati dobivene modele klasifikacije na neviđenim podacima iz dostupnog skupa. Implementirati odabrane algoritme za objašnjivost predikcija modela strojnog učenja, te analizirati rezultate. Za detaljne informacije obratiti se Edi Jovičić, mag. ing.

Rok za predaju rada: 28. lipnja 2024.

I would like to start by expressing my gratitude to my thesis mentor, Professor Mario Cifrek, for taking me as a mentee. As a person driven by purpose rather than process, it was important for me to finish my degree working on a topic I truly believe in, and I'm incredibly grateful for the opportunity to work on this highly interesting and relevant topic.

My former mentor, Associate Professor Jurica Babić, also deserves a special mention. Although we had to part ways in terms of our scientific endeavors, I continue to use the knowledge and motivation he provided me with, and I am sure I will use it in the future.

Another big thanks to Eda Jovičić for preparing the data for this thesis and providing valuable support in the scary landscape of signal processing.

I would also like to extend my gratitude to the Psychiatric Hospital Vrapče and all the study participants involved in the collection of the dataset. This research would not have been possible without their contributions and willingness to participate.

I can't leave FER without thanking its Mobility Office, specifically Ljiljana Brkić and Neda Tomaš, whose work heavily contributed to my immense growth experience of a year abroad at Technische Universität Wien.

Thank you to my parents for everything they provided me with and for shaping me into the person I am today. Writing a thesis is one thing but raising a human is another. I can never repay all they've given me, but I hope they find satisfaction in what I am able to achieve thanks to their efforts.

I also want to thank my sister, friends, and colleagues for their support both in my studies and personal life. As we move forward in our careers and lives, I hope we always cherish these shared moments.

Finally, I would like to acknowledge that I wouldn't have been able to complete this thesis without overcoming personal challenges. I'm proud to say that as I am finishing my degree, I am also finishing my antidepressant therapy. Though my treatment was not related to this diagnosis, I dedicate this work to all those suffering from depression, hoping that my research, however small, contributes to understanding and addressing of your pain.

And what I can offer as comfort from personal experience:

No feeling is final.
– R.M. Rilke

Contents

1 Introduction

Depression, which is becoming an increasingly prevalent mental health disorder in our fast-paced modern society, represents a complex and multifaceted challenge for both individuals and society as a whole. It has the potential to impact all aspects of a person's life, from their ability to perform daily tasks to their long-term mental and physical well-being. It can influence their relationships, their work, their self-esteem and even their outlook on life. Thus, depression should not be taken lightly and requires comprehensive understanding and careful management.

Current diagnostic methods for depression often rely heavily on subjective assessments, such as self-reported symptoms and clinical interviews. While these methods provide valuable information, they are inherently prone to variability due to differences in interpretation, reporting bias, and the complexity of depressive symptoms. This unreliability can cause inconsistent diagnoses and delay proper treatment. It can also lead to misinterpretations of the range and intensity of depression symptoms. People might not report all their symptoms due to fear of being judged or not being aware of their mental health status. On the other hand, symptoms might be reported excessively, resulting in wrong diagnosis or unnecessary treatment. Additionally, these subjective evaluations might not capture depression's neurobiological aspects completely. Depression is increasingly recognized as a disorder with a range of neurobiological factors, like changes in brain structure and operation. Objective measurements would offer a way to directly assess these neurobiological markers, providing insights into brain activity patterns that correlate with depressive states.

In recent years, the use of EEG brain recordings has emerged as a promising avenue for this objective measurement of depression. EEG captures electrical activity in the brain non-invasively, offering objective insights into neural patterns associated with depres-

sion while also not requiring immense procedures. Other than it holding potential for improving diagnostic precision it also shows potential for guiding targeted therapeutic strategies. By understanding the specific neural patterns associated with each individual's depression, treatments could be personalized to address the unique needs of each patient. Overall, the use of EEG in diagnosing and treating depression holds significant potential for revolutionizing mental healthcare.

Significant strides have already been made in utilizing EEG as a biomarker for depression. Various machine learning models have been employed to classify individuals as healthy or depressed based on EEG data. These models have demonstrated promising performance, suggesting EEG's potential as a reliable biomarker for depression. This raises the question: what specific characteristics of the EEG signal contribute to its effective prediction of depression? This thesis aims to investigate this question through the lens of explainable AI (XAI), an emerging field focused on understanding the performance of complex machine learning models.

1.1 Objectives

The objectives of this thesis are twofold:

Identifying Potential Biomarkers of Depression serves as the primary objective and motivation for this thesis. The aim is to uncover certain EEG signal characteristics that reliably predict the presence of depression.

Comparison of XAI Methods is the secondary objective of the thesis, aiming to identify the advantages and disadvantages of different XAI methods and their applicability in understanding biomarkers of depression.

By addressing these objectives, this thesis aims to contribute to the fields of mental health and data science, offering both insights into biomarkers of depression and practical insights into the use of XAI methods.

1.2 Structure

The rest of this thesis is divided into seven main chapters. Chapter 2: Key Concepts briefly introduces affective disorders, electroencephalography (EEG), machine learning and explainable artificial intelligence (XAI). Chapter 3: Dataset details the main characteristics, collection and preparation procedures, as well as prior research applications of the used dataset. Chapter 4: Literature Review presents a comprehensive literature review focusing on XAI in EEG applications and its predictive potential for affective disorders. Chapter 5: Methodology encompasses experiment design, setup, chosen and rejected methods, and limitations of this thesis. Chapter 6: Results lists all relevant outcomes of analyses. Chapter 7: Discussion addresses possible reasons for achieved results, and lastly, Chapter 8: Conclusion provides a summary of the thesis. At the end, references, abstracts, and appendices are provided.

2 Key Concepts

This chapter establishes the foundations of this thesis, providing basic definitions and background information necessary for comprehending the complexities of the topic.

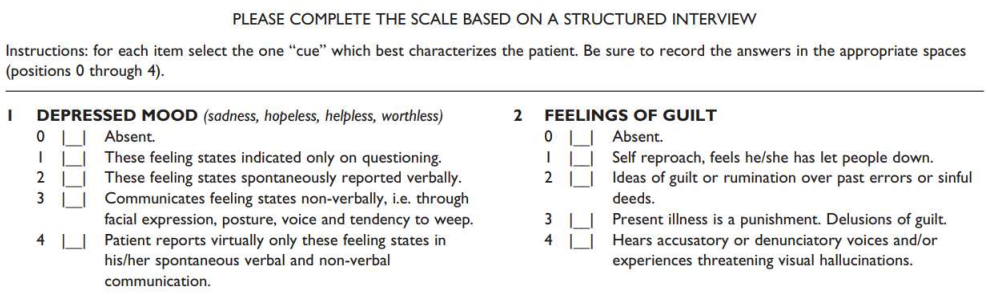
2.1 Affective Disorders

Affective disorders are considered a subgroup of mental and behavioral disorders characterized by a shift in mood to either elation or depression, with or without anxiety [1]. Typically, this mood shift is also accompanied by a change in general activity level. Disorders of this group are usually recurring and associated with stressful events or situations.

These disorders can roughly be further divided into three main categories: depressive disorders, characterized by persistent low mood; manic disorders marked by unusually elevated mood; and bipolar disorders, which involve fluctuations between depressive lows and manic highs.

2.1.1 Diagnosis

The traditional approach to diagnosing affective disorders, which remains prevalent, involves the use of non-laboratory techniques such as patient interviews and questionnaires. An example of such a questionnaire is the *Hamilton Depression Rating Scale (HDRS or HAM-D)* represented in figure 2.1. In such diagnostic processes, clinicians typically seek out signs and symptoms of affective disorders as outlined in recognized classification manuals, notably the *Diagnostic and Statistical Manual of Mental Disorders (DSM)* issued by the *American Psychiatric Association (APA)* and the *International Classification of Diseases (ICD)* issued by the *World Health Organization (WHO)*. Over the years both manuals have undergone significant changes in response to scientific ad-



PLEASE COMPLETE THE SCALE BASED ON A STRUCTURED INTERVIEW

Instructions: for each item select the one "cue" which best characterizes the patient. Be sure to record the answers in the appropriate spaces (positions 0 through 4).

1 **DEPRESSED MOOD** *(sadness, hopeless, helpless, worthless)*
- 0 |__| Absent.
- 1 |__| These feeling states indicated only on questioning.
- 2 |__| These feeling states spontaneously reported verbally.
- 3 |__| Communicates feeling states non-verbally, i.e. through facial expression, posture, voice and tendency to weep.
- 4 |__| Patient reports virtually only these feeling states in his/her spontaneous verbal and non-verbal communication.

2 **FEELINGS OF GUILT**
- 0 |__| Absent.
- 1 |__| Self reproach, feels he/she has let people down.
- 2 |__| Ideas of guilt or rumination over past errors or sinful deeds.
- 3 |__| Present illness is a punishment. Delusions of guilt.
- 4 |__| Hears accusatory or denunciatory voices and/or experiences threatening visual hallucinations.

Figure 2.1: Hamilton Depression Rating Scale [3]

vancements, resulting in multiple editions, with the latest iterations being DSM-V and ICD-11.

These non-laboratory techniques suffer from various problems leading to inconsistent and sometimes incorrect diagnoses, most commonly; the subjective nature of interpretation, limited patient information and reporting, and the presence of comorbidity with overlapping symptoms making it difficult to pinpoint the primary diagnosis [2].

2.1.2 Depression

Outside of clinical contexts, when "depression" is colloquially mentioned, it is usually referred to a diagnosis known as a depressive episode, or if it happens repeatedly, recurrent depressive disorder.

A **depressive episode** is characterized by a period of lowered mood, energy, and activity lasting at least two weeks, during which symptoms are present most of the day, nearly every day [1]. It may also be accompanied by the loss of interest, enjoyment, appetite, libido, and the ability to concentrate. Feelings of hopelessness and worthlessness tend to also occur. In ICD-10 a depressive disorder is identified by the **F32** code, with the following subtypes:

F32.0 Mild depressive episode
F32.1 Moderate depressive episode
F32.2 Severe depressive episode without psychotic symptoms
F32.3 Severe depressive episode with psychotic symptoms
F32.8 Other depressive episodes
F32.9 Depressive episode, unspecified

The **reccurent depressive disorder** is defined by a history of at least two depressive episodes separated by at least several months without significant mood disturbance [4]. Additionally, to receive this diagnosis, individuals must meet specific criteria outlined in the diagnostic manuals. Notably, there should be no prior manic, hypomanic, or mixed episodes, which would indicate the presence of bipolar disorder. In ICD-10 recurrent depressive disorder is identified by the **F33** code, with the following subtypes:

F33.0 Recurrent depressive disorder, current episode mild
F33.1 Recurrent depressive disorder, current episode moderate
F33.2 Recurrent depressive disorder, current episode severe without psychotic symptoms
F33.3 Recurrent depressive disorder, current episode severe with psychotic symptoms
F33.4 Recurrent depressive disorder, currently in remission
F33.8 Other recurrent depressive disorders
F33.9 Recurrent depressive disorder, unspecified

For the sake of simplicity in this thesis, any diagnosis falling under any of the categories **F32.x** or **F33.x** will be referred to as "depression".

2.2 Electroencephalography (EEG)

Electroencephalography, often abbreviated as **EEG**, is a method used to record and analyze the electrical activity of the brain, first introduced in 1929 by German scientist Hans Berger [5]. Pairs of electrodes are attached to the scalp to detect rhythmic electrical signals generated by the brain's neurons. Each electrode pair relays voltage differences visualized as line graphs on the electroencephalogram. An example of an EEG recording is shown in figure 2.2.

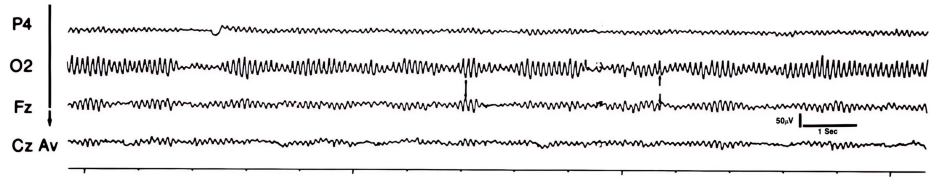


Figure 2.2: Example of an EEG reading [6]

2.2.1 Physiological Basis

The brain consists of billions of neurons exchanging information by electrical and chemical signals. Simplified representation of a neuron is shown in figure 2.3, to visually enhance the understanding of the complex processes described. Brief changes in the membrane potential of neurons are called action potentials, and they are propagated along neurons' axons, triggering neurotransmitter release at synapses. These neurotransmitters then excite or inhibit neighboring neurons, enabling signal propagation through multiple neurons across the human brain and body.

Neurons in the brain often synchronize their activity to perform various functions, such as processing sensory information, generating movements, or producing thoughts. When groups of neurons synchronize their activity, they generate rhythmic patterns of electrical activity known as brain oscillations. These oscillations occur at different frequencies and are associated with different cognitive processes and states of consciousness.

When the electrical activity of large groups of neurons sums up, its activity can be detected in nearby areas, such as the scalp. The electroencephalography method of brain imaging utilizes this fact and captures the electrical activity with electrodes placed on the scalp. These electrodes measure the voltage fluctuations resulting from the collective electrical activity of neurons firing synchronously within specific brain regions or networks.

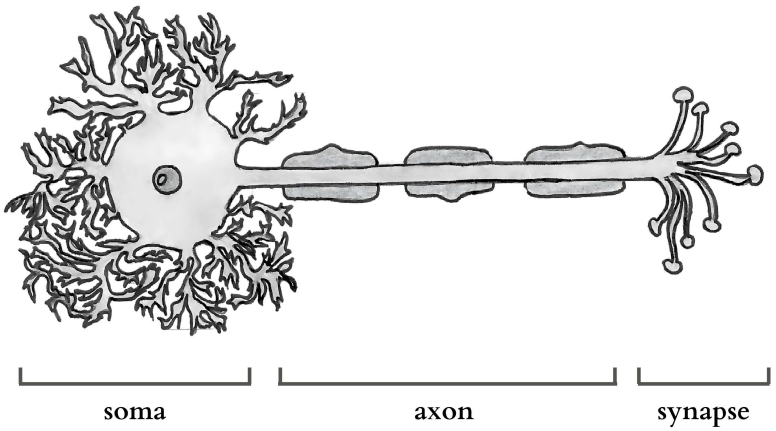


Figure 2.3: Schema of the neuron showing the major structural features

2.2.2 Recording

The equipment required for recording a modern EEG is simple and inexpensive compared to other brain imaging methods [7]. The method is noninvasive and requires only a set of electrodes, a signal amplifier, and a computer. A simple representation of the recording process can be seen in figure 2.4.

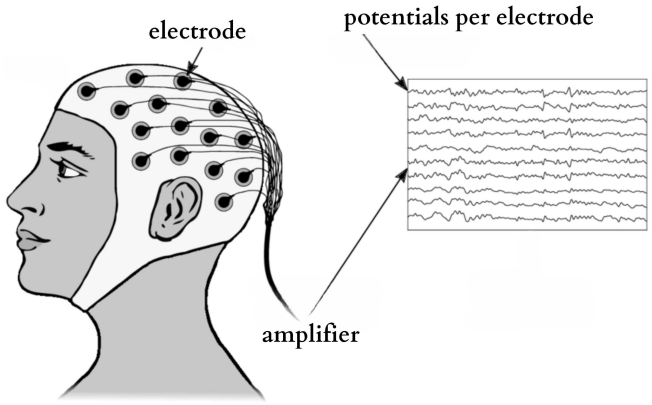


Figure 2.4: Visual representation of the EEG recording process and result [8]

EEG electrodes can be arranged in various configurations, varying in the number and position of electrodes along the scalp. To enable comparison between the work of different laboratories and studies, the First International Congress of EEG in 1947 recommended standardizing the electrode configuration. The **10-20 International System** was soon proposed and it is still the most used configuration today [9]. The name of the configuration stems from electrode positions spaced at intervals of either 10% or 20% of the total front-to-back distance of the skull, more specifically, the distance between the nasion and the inion. The use of percentages instead of absolute values for electrode placement accounts for individual variances in head size and shape. For an easier understanding of the system, figure 2.5 can be referenced. Each electrode is identified by a label consisting of a letter and a number, for example, Fp2. Letters indicate the corresponding area of the brain the electrode is reading from; Fp for prefrontal, F for frontal, C for central, T for temporal, P for parietal, O for occipital, and A for earlobes (mastoid). Odd numbers denote electrodes situated over the left hemisphere, while even numbers denote those over the right hemisphere of the brain. Additionally, there are special electrodes placed along the line from nasion to inion, marked with "z" instead of a number.

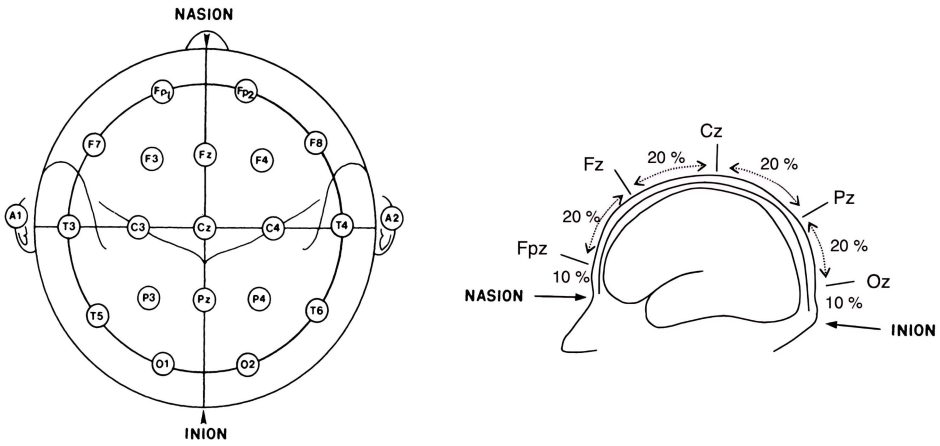


Figure 2.5: The 10-20 International System for electrode placement [6, 10]

When recording a more detailed EEG, instead of 10-20, a 10-10 system can be used, which utilizes more electrodes, placed between the electrodes of the previously explained 10-20 system following the same logic.

2.2.3 EEG Brain Waves

Different frequency ranges and amplitudes distinguish five major EEG brain waves [6, 11] marked by Greek letters: delta (δ), theta (θ), alpha (α), beta (β), and gamma (γ) represented in 2.6.

Delta (0.5–4 Hz) waves are primarily associated with deep sleep. They have the highest amplitude of all waves and are thought to play a role in transferring learning and long-term memory storage.

Theta (4–7.5 Hz) waves are observed during daydreaming or light sleep and are linked to accessing unconscious material, creative inspiration, and deep meditation. They are more prevalent in children and adolescents compared to adults and have been linked to coordinating the process of memory storage.

Alpha (8–13 Hz) is the wave normally recorded in the awake relaxed individual with eyes closed. It occurs dominantly in the posterior half of the head and appears round or sinusoidal-shaped.

Beta (14-26 Hz) waves are associated with active thinking, attention, and focus. They are most prominent in the parietal and frontal regions of the scalp, as this is where

the abovementioned mental tasks take place.

Gamma (30+ Hz) occurs rarely and can be used for confirmation of certain brain diseases. It is also associated with a stressed brain under heavy workload. The resolution of EEG does not allow the naked eye to examine this frequency range in a clinically useful manner [10], thus it is omitted from the visualization below.

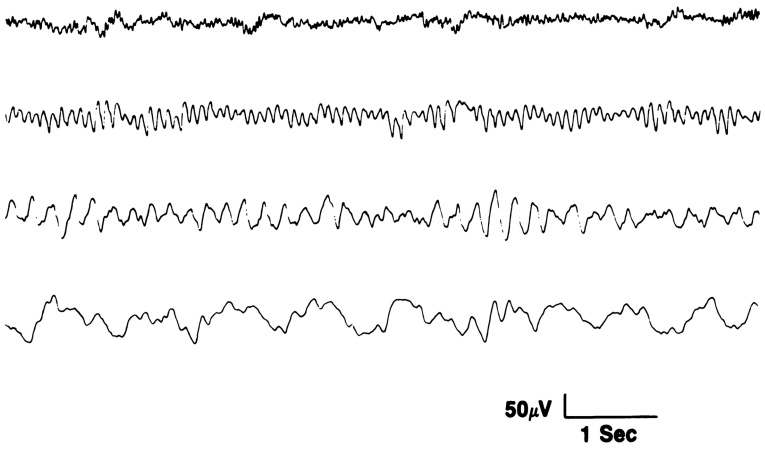


Figure 2.6: Signature brain wave patterns, from top to bottom: beta, alpha, theta, delta [6]

2.2.4 EEG Preprocessing

Preprocessing refers to a series of signal-processing procedures conducted before the main analysis of EEG data. Because of the inherent noise in EEG recordings, it is a necessary step in any EEG study. Noise can be caused by internal factors, notably eye and muscle movements or pulse, and external factors, such as power line interference and movement of electrodes or cables. These factors are commonly referred to as artifacts, and EEG preprocessing is mainly focused on removing artifact effects from the EEG signals. There is no standardized EEG preprocessing pipeline, but a short overview of common methods based on [12, 13] is given below.

Filtering is an essential preprocessing step in EEG signal processing. EEG signals contain noise that can distort or obscure meaningful brain activity. For high-frequency noises from sources such as muscle activity or environmental electromagnetic interference, a low-pass filter can be used. For low-frequency noises from sources like electrode drift or movement artifacts, a high-pass filter can be used. Band-pass filters can be used for filtering both high- and low-frequency noises simultaneously.

Band-stop (also known as notch) filters can be utilized for removing power line interference at 50 Hz in Europe and Asia or at 60 Hz in the United States.

Referencing is another method of removing the noise by subtracting a reference signal from the original EEG signal at each channel. By comparing the activity recorded at the scalp electrodes to a stable reference point, it's possible to reduce the effects of noise unrelated to brain function. Typical choices of reference are the mastoid channel, the average of two mastoid signals, or the average of all channels.

Bad channel removal and interpolation is another important step of preprocessing. Some channels may be improperly placed on the scalp making the recorded EEG signal corrupted and unusable. They are referred to as "bad channels" and should be removed from analysis. However, to not reduce comparability among individuals with varying bad channels, the bad channels are usually interpolated from the "good channels", most commonly using spherical splines, higher-order polynomials, nearest-neighbor averaging, or radial basis function.

Artifact removal methods of preprocessing usually revolve around the blind source separation (BSS) principle. BSS aims to decompose a set of linearly mixed signals into a set of distinct source signals with "little knowledge" about the source signals or the mixing process, ultimately separating the brain signals from the noise. Different methods based on BSS differ in the definition of "little knowledge" in terms of assumptions on the source signals. Most commonly utilized, independent component analysis (ICA) assumes linear independence and non-Gaussianity, and canonical correlation analysis (CCA) assumes maximal uncorrelatedness.

2.2.5 EEG Feature Extraction

EEG feature extraction is the process of extracting relevant information from EEG signals. Typically, EEG signals contain a large amount of data, because of multiple electrodes, long recording procedures, and a high temporal resolution of millisecond precision. Extracting features from raw EEG data effectively reduces the volume of data required for subsequent analyses while retaining important information. This also allows for the application of various machine learning algorithms because it converts time-series EEG data into tabular data, with each row representing a sample (e.g., an individual's recording session) and each column representing a specific feature.

Features can be extracted through various domains [14]:

Time domain features are typically the least complex to extract and include statistical measures (mean, variance, skewness...), autoregression, and fractal dimension.

Frequency domain features are extracted from the sinusoids that make up the data, after the conversion of signals from time to frequency domain. Typical features from this domain include power spectral density, absolute band power and relative band power.

Time-frequency domain overcomes the limitations of time and frequency domains alone by combining them. Common features include those based on the short-time Fourier transform (power, entropy, centroid), wavelet transform (coefficients, entropy, energy), and Hilbert-Huang transform (instantaneous frequency, spectrum).

Spatial domain methods convert the brain waves into a unique space, where the variance of one group is magnified, and a lower variance is seen in the remaining group. The most commonly used method is common spatial pattern (CSP).

Some example features are further introduced below for context in further analyses of this thesis.

Absolute Band Power represents the total power contained within a specific signal frequency band.

Relative Band Power is the proportion of power within a specific frequency band relative to the total power across all frequency bands.

Spectral Centroids indicates the center of mass of the spectrum of a signal, providing a sense of where most of the energy in the signal is concentrated in terms of frequency. A higher value typically indicates that the signal has more high-frequency components, whereas a lower spectral centroid suggests lower frequencies.

Relative Wavelet Energy is the proportion of energy within a specific frequency band relative to the total energy across all frequency bands, derived from the wavelet transform.

Wavelet Entropy measures the spectral complexity of a signal in terms of its energy distribution across different frequency components. The concept of entropy stems from information theory and thus reflects the information content of the signal's frequency components.

Katz Fractal Dimension quantifies the complexity or roughness of a waveform, similar to wavelet entropy but through a geometry lens instead of information theory. A higher value indicates a more complex signal and a lower value suggests a smoother and simpler signal.

2.3 Machine Learning

Machine learning uses example data to "learn" the necessary transformation algorithm from input to output or from input to extracted information. More formally, machine learning includes programming computers to optimize a performance criterion using example data or past experience [15]. This can involve supervised learning, where the example data includes both inputs and desired outputs, and the performance criterion is most commonly the proportion of correctly guessed examples, or unsupervised learning, where the goal is to discover inner structures or patterns within the data, for example, grouping, with one possible performance criterion being the similarity of an example with its own group compared to other groups. In this thesis, further focus will be on supervised learning, as the available dataset contains EEG data as inputs and diagnoses as outputs for the machine learning problem. More specifically, this is a classification problem, meaning the outputs are classes: healthy individuals or patients with diagnosed depression.

In the supervised learning process of machine learning, the objective is to find an effective mapping function from input to output data. This mapping function, called a **hypothesis**, is defined by a set of **parameters**. The search for a good hypothesis is done within a **model** – a predefined set of possible hypotheses – and involves an **optimization process** of adjusting the parameters to explore different hypotheses within the model, ultimately aiming to find the hypothesis that minimizes the **loss** – the difference between the desired outputs and the outputs predicted by the hypothesis.

Once the "best" hypothesis is found, it is crucial to check how well it generalizes to new, unseen data. For this purpose, a **test set** is usually prepared - a portion of example data is put aside not to be used in the optimization. The example data used in the optimization process is called the **training set**.

Ideally, there exists a hypothesis within the model for which the loss on the training set is zero, however, in most real cases this is impossible due to the presence of **noise**. Real-life datasets usually contain some incorrect pairs of input and output, such as an image of a tumor falsely labeled as healthy. A hypothesis that fits this kind of wrongly labeled data too closely would be overly complex and fail to generalize well to new examples. If a model is too simple, it may not capture the underlying patterns in the data, referred to as the problem of **underfitting**. Conversely, if a model is too complex, it may lead to the problem of **overfitting**, capturing noise instead of the actual data patterns. The complexity of a model is controlled by the model's **hyperparameters**, which are higher-level structural settings. Hyperparameters are not learned from the example data like parameters of a hypothesis; instead, they are set before the learning process.

Ultimately, the goal is to find a model (more specifically, a hypothesis within a model) that not only minimizes loss on the training data but also generalizes well to test data, achieving a **balance between underfitting and overfitting**. This notion is in some literature also referred to as the **bias-variance trade-off**, represented in figure 2.7. Because overfitted model fits the noise too well, instead of the underlying trends in the data, it is too sensitive to fluctuations and results in **high variance**, performing well on train examples but poorly on test examples. On the other hand, an underfitted model pays little attention to the data and oversimplifies the mapping from input to output, resulting in **high bias** and poor performance on both the train and test examples.

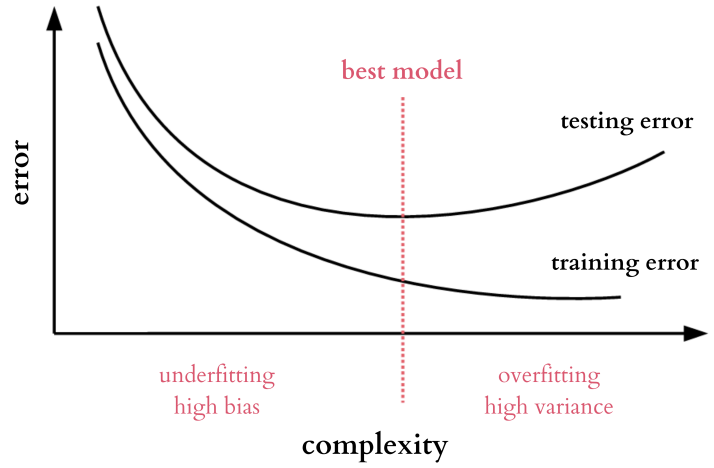


Figure 2.7: Bias-variance trade-off illustrated, adapted from [16]

2.3.1 Model Evaluation

For supervised classification problems, the most commonly used evaluation measures include accuracy, precision, recall, and F1-score. All of these metrics stem from the **confusion (contingency) matrix** represented below in 2.1. The confusion matrix is a square matrix that summarizes the number of matches and mismatches between predicted and desired (true) outputs. For a binary classification problem, it consists of the following four elements:

- true positives (TP) – predicted and desired output is 1
- true negatives (TN) – predicted and desired output is 0
- false positives (FP) – predicted output is 1 and desired output is 0
- false negatives (FN) – predicted output is 0 and desired output is 1

$$\begin{array}{c} \\ y_{pred}=1 \\ y_{pred}=0 \end{array} \begin{array}{c} \begin{array}{cc} y_{true}=1 & y_{true}=0 \end{array} \\ \begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix} \end{array} \tag{2.1}$$

The most obvious and commonly used metric derived from the confusion matrix is **accuracy**, the proportion of correctly classified examples given by 2.2.

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \tag{2.2}$$

However, accuracy suffers when an unbalanced dataset is used, meaning one class is represented more than another, and a model that would always guess the majority class would already have a higher than 50% accuracy. This is why further metrics have been developed, most notably precision and recall.

Precision is the proportion of correct positive identifications, given by 2.3.

$$P = \frac{TP}{TP + FP} \tag{2.3}$$

Recall is the proportion of correctly identified actual positives, given by 2.4

$$R = \frac{TP}{TP + FN} \tag{2.4}$$

Precision and recall are considered to be opposite in the sense that improving one can often lead to a reduction of the other. A measure that combines precision and recall, and thus gives a holistic measurement of the performance of a model is the **F1-score**. Formally, it is the harmonic mean of precision and recall, as shown in 2.5.

$$F1 = \frac{2}{\frac{1}{P} + \frac{1}{R}} \tag{2.5}$$

In some applications, the generalized $\mathbf{F}_\beta$**-score** is preferred, given by 2.6, which provides a weight β adjustment representing how many times recall is as important as precision. With $\beta = 1$ it reduces to the previously introduced F1-score.

$$F_\beta = \frac{1 + \beta^2}{\frac{1}{P} + \frac{\beta^2}{R}} \tag{2.6}$$

2.4 Explainable Artificial Intelligence (XAI)

The field of explainable AI (XAI) seeks to address challenges that arose recently as a result of big advancements in machine learning which are making models so complex that humans do not easily understand them. It is focused on providing methods that make AI decision-making processes more transparent and understandable. This is particularly important in the fields of healthcare, finance, and criminal justice, where the rationale behind decisions must be clear and accountable.

The field is particularly interesting because it attempts to mitigate the trade-off between model accuracy and interpretability. Very accurate models tend to be the least easy to understand for humans, as shown by the graph in figure 2.8.

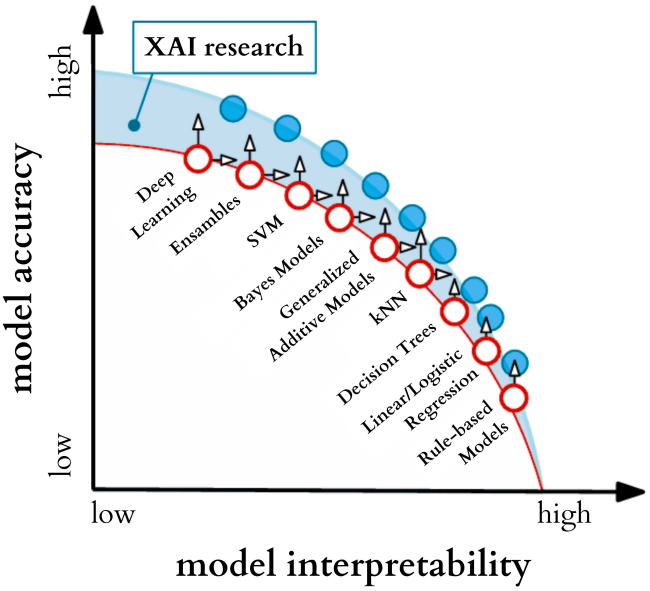


Figure 2.8: Model performance and interpretability trade-off illustrated, adapted from [17]

2.4.1 Terminology

The definitions behind what makes AI understandable to humans are unfortunately still ill-defined. Various keywords such as "transparency", "interpretability", "intelligibility", "comprehensibility" and others, as extensively outlined in [18], are used in a somewhat interchangeable manner in literature without a unified vision and understanding of the terms [19]. Some efforts have been made to settle on definitions, one significant being the ISO [20], but widespread agreement on this terminology doesn't seem to be achieved in academia or media:

Interpretability (3.1.42) – level of understanding how the underlying (AI) technology works

Explainability (3.1.31) – level of understanding how the AI-based system came up with a given result

2.4.2 Purposes

The purposes of XAI are numerous and suffer from a similar problem with terminology as the field itself. For the purpose of this thesis, a categorization inspired by [21] and adjusted to [20] is provided. This overview is not intended to be exhaustive or mutually exclusive but to provide a comprehensive understanding.

XAI purposes can be categorized into two main areas: understanding the inner workings of an AI system (interpretability) and understanding the outputs of an AI system (explainability):

Interpretability-focused purposes can be seen as more closely related to scientific purposes of informativeness, transferability, and causality:

- **Informativeness.** By understanding how models process data and make decisions more deeply, researchers can identify areas for improvement, leading to the enhancement of future AI models. This depth of understanding enables the detection of flaws and biases, fostering development of more robust and accurate AI systems.
- **Transferability.** When researchers thoroughly understand an AI model, they can adapt it to new problems and datasets, enhancing the model's versatility and utility.
- **Causality.** At the core of scientific endeavors lies proving causality, which typically requires significant domain knowledge and extensive experimentation. However, XAI shows promise in identifying cause-and-effect relationships [21].

Explainability-focused purposes are more closely related to practical applications of AI systems, where fairness and bias, accountability, and safety play a big role.

- **Fairness and bias.** XAI aims to identify and mitigate unfair or unethical biases embedded within AI systems. These biases often arise from training data that reflects societal prejudices, such as gender and racial biases.
- **Accountability.** In many applications accountability is important due to legal and regulatory frameworks that require justifications for decision-making processes. XAI works on providing transparent reasoning behind AI decisions, ensuring that stakeholders can understand the outcomes.
- **Safety.** Understanding the behavior of AI systems is essential for ensuring safety in high-stakes applications. In domains such as autonomous driving and aviation, the ability to explain AI decision-making processes can prevent accidents and save lives.

Ultimately, the crown purposes of XAI are **trust and adoption** of AI in society, for which both interpretability and explainability play an important role. Without trust, people hesitate to rely on and use AI systems, and without adoption, AI stays in laboratories and papers, missing out on its potential to be applied in industries and improve lives.

2.4.3 Taxonomy

As with anything related to XAI, there have also been numerous nonunified taxonomies offered for categorizing XAI methods. According to [21], represented in figure 2.9, the three main divisions appear to be based on stage, scope, and model.

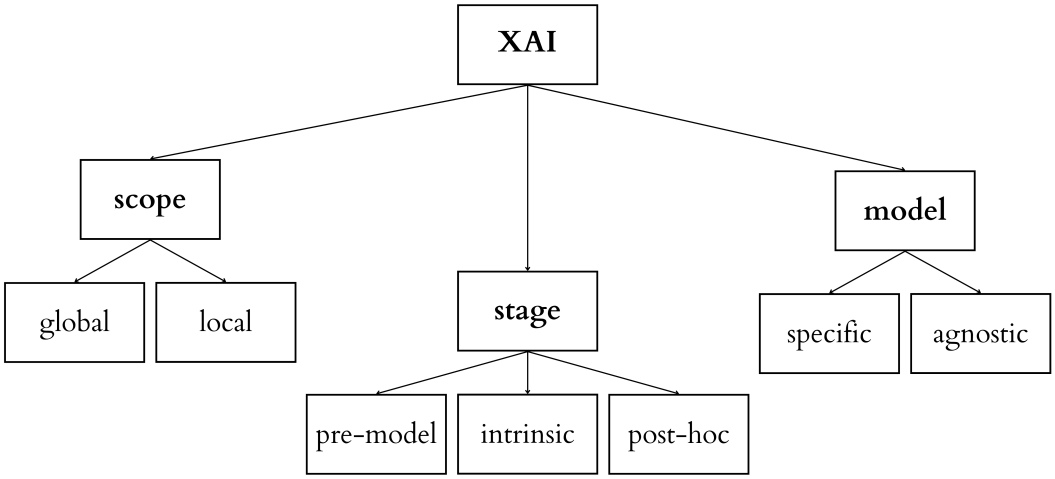
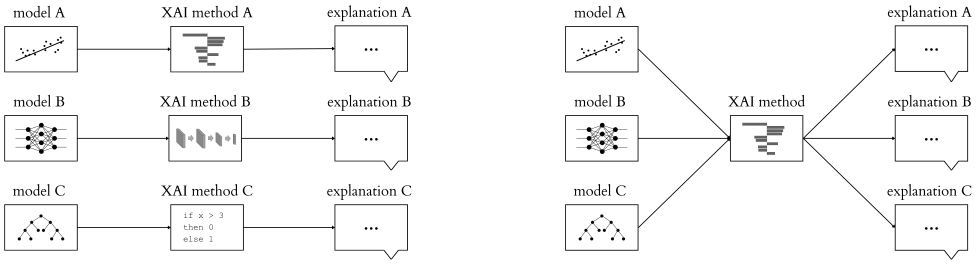


Figure 2.9: Taxonomy of XAI, adapted from [21]

Model

This categorization takes into account whether the XAI method applies to only specific models or a wider range of them. **Model-agnostic** methods work on any model. These include both pre-model and post-hoc methods from stage taxonomy. **Model-specific methods** can only be applied to some models. They usually offer more specific and detailed explanations that model-agnostic methods can't provide.



(a) Model-specific (b) Model-agnostic

Figure 2.10: Model Taxonomy of XAI illustrated, adapted from [22]

Scope

The scope divides the XAI methods into global and local. **Global** methods aim to provide a holistic comprehension of the entire model. **Local** methods focus on explaining the decisions of a model on a single example.

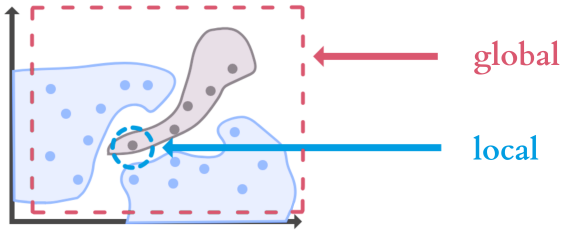


Figure 2.11: Scope Taxonomy of XAI illustrated, adapted from [23]

Stage

This taxonomy categorizes XAI methods based on at which stage of the AI pipeline they operate; before, during, or after a model makes its prediction.

Pre-Model methods are applied directly to the data before model training, independent of the model itself. They involve various exploratory data analyses and data visualizations. Even if interpretability or explainability are not explicit goals of an AI project these methods are usually performed because they help data scientists develop an understanding of the data and the problem at hand.

Intrinsic XAI methods include using naturally easily interpretable machine learning models such as decision trees, naive Bayes, k-nearest neighbors, or rule-based models. These are also sometimes named "white box" algorithms, because of their inherent transparency.

Post-hoc methods are applied at the end of the pipeline, and just like pre-model methods, are independent of the model itself, they can be used on any model. They provide explanations based on the relationship between the input and output of the model. They are especially useful for explaining complex models such as neural networks or transformers, which are often named "black box" models because their internal workings are not easily understandable.

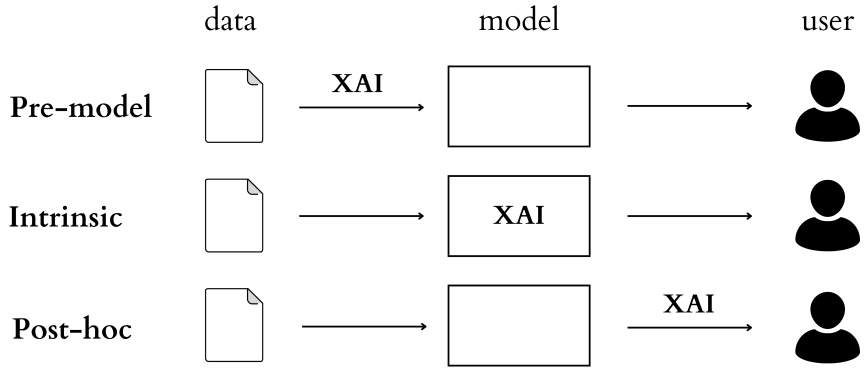


Figure 2.12: Stage Taxonomy of XAI illustrated, adapted from [21]

2.4.4 XAI Evaluation

Despite the lack of consensus in defining explainability terms, purposes and taxonomies, there have also been initial efforts made on how to measure it. The evaluation methods can be separated into three approaches [24]. **The application-grounded approach** involves putting the explanation directly into the system for the target end user. This approach is compliant with the ethos of the human-computer interaction field, where a large focus is on whether the system delivers on its intended task. **The human-grounded approach** is based on the same idea as the application-grounded approach, but is easier to achieve. It involves simpler human-subject experiments that keep the essence of the

target application. This is especially useful for applications for which the target end user is a domain expert, for example in the medical domain, and it is expensive to conduct special experiments. Lastly, the **functionally-grounded approach** doesn't require any humans to evaluate the AI; instead, it uses some formal definition of interpretability as a proxy. For example, the depth of the tree could be the proxy for the explanation quality of decision trees, and shorter trees would get a better explainability score.

3 Dataset

This chapter serves to introduce the dataset utilized in this thesis. Mainly, the data collection and preparation are presented and the final dataset is described for better understanding in subsequent analyses.

3.1 Description

The dataset was collected at the Psychiatric Hospital Vrapče in Zagreb and contains EEG recordings of a total of **105 individuals** recorded over the period from 2016 to 2020. To ensure balanced representation, the dataset was structured to include **70 patients** with affective disorders and **35 healthy** individuals. Each healthy individual was matched with two patients with an affective disorder of the same sex and similar age. The patients with affective disorders were diagnosed according to the ICD-10 [1], as the dataset was developed before the adoption of the ICD-11 [4]. For machine learning purposes, the data is further divided into a **training set of 75 subjects** and a **testing set of 30 subjects**.

The data collection was performed under the guidelines of the Declaration of Helsinki and approved by the ethics committee of the University of Zagreb.

3.2 Recording procedure

The EEG was recorded at a sampling frequency of 256 Hz using the 10-20 system with 19 electrodes including Fp1, Fp2, F7, F3, Fz, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1, O2, and Oz as the reference electrode.

The recording procedure consisted of three activities:

1. **Resting** with eyes opened and closed (5-10 minutes),
2. **Photo stimulation** with five different flash frequencies (15 seconds each),
3. **Induced hyperventilation** (5 minutes),

accompanied by manual marking of each event onset by a technician.

3.3 Preprocessing and Feature Extraction Procedures

After recording, the data was further prepared with preprocessing and feature extraction at the Faculty of Electrical Engineering and Computing in Zagreb, as shown in 3.1 [25].

Preprocessing included band-pass filtering with a passband ranging from 0.1 to 40 Hz, referencing with the average of all channels as the reference, and artifact removal with ICA. For feature extraction, the signal was first decomposed into frequency bands corresponding to characteristic brain rhythms with discrete wavelet transform (DWT). Subsequently, six features were extracted for each of the five main EEG brain rhythms (delta, theta, alpha, beta, and gamma): absolute band power, relative band power, spectral centroids, relative wavelet energy, wavelet entropy, and Katz fractal dimension.

With a total of 19 electrodes, 6 features, and 5 characteristic brain waves, this amounts to a total of **570 attributes** per individual.

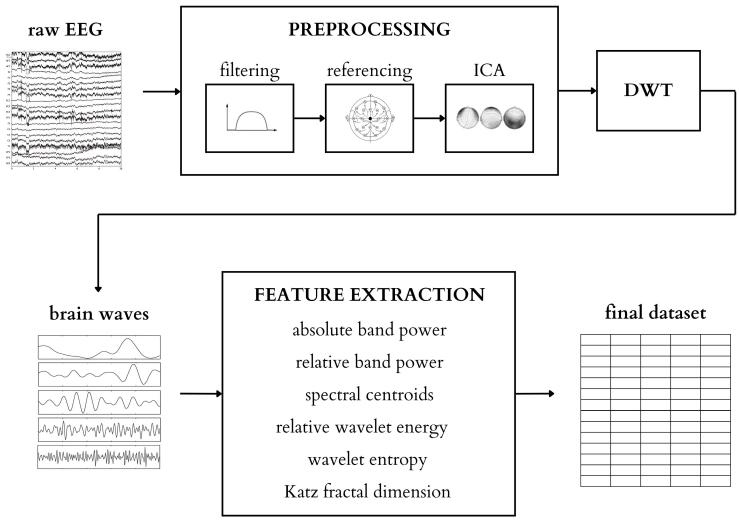


Figure 3.1: Dataset preparation process

3.4 Prior Research on the Dataset

The dataset has been used for scientific research in several other works.

- Different machine learning methods have been tried for the classification of affective disorders in [26], including SVM, Naive Bayes, Linear Discriminant, Tree, kNN, and Logistic Regression, with results showing the highest accuracy and F1 score for SVM and kNN models.

- Different preprocessing methods have been tried and tested in terms of their effect on the performance of machine learning models in [25], including filtering, canonical correlation analysis, and independent component analysis, with results showing canonical correlation analysis as the best preprocessing technique.

- Lastly, major depressive disorder in patients with borderline personality disorder was investigated in [27]. The study showed patients suffering from major depressive disorder did not have significantly different EEGs from patients diagnosed with major depressive disorder and borderline personality disorder.

4 Literature Review

After introducing the key concepts and the methodological framework of explainable AI, this chapter aims to provide an overview of previously performed scientific endeavors of XAI on EEG data and XAI in predicting affective disorders, as well as their intersection.

4.1 XAI in EEG Applications

While numerous studies highlight the theoretical potential and significance of XAI in medical EEG applications [28, 29], the primary focus in the available literature is on deep models predicting specific medical disorders. This rise of XAI to explain deep learning in EEG is illustrated by the graph in figure 4.1.

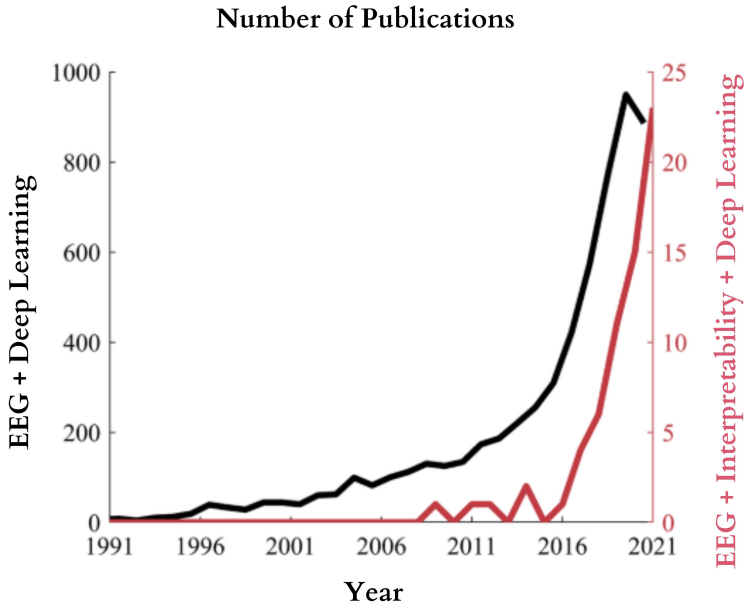


Figure 4.1: Trend showing the number of EEG publications using deep learning, with and without explainability, adapted from [30]

4.2 XAI in Prediction of Affective Disorders

For the prediction of affective disorders, leveraging textual data from social media has become one of the most commonly used methods, as seen in studies like [31, 32]. This trend is unsurprising with the vast amount of textual data available and the advancements in the field of Natural Language Processing (NLP). Naturally, with the increasing complexity of NLP models, the need for XAI has become paramount.

Some work has focused on mental health apps instead of social media. These apps offer functionalities such as mood tracking, predicting affective disorders, and delivering mental health advice. While they hold great potential in reaching a wider audience and providing timely support, there are concerns about the lack of AI literacy among these apps. A review of 13 AI-based mental health apps suggests the need for employment of XAI [33].

Lastly, some work focused on the usage of other possible biomarkers like diet, heart rate and heart rate variability, physical activity, sleep, breathing and neurological task performance. In the example of [34], these were used to predict mood with various decision tree based models, ensuring inherent explainability. Synthesis of explanations revealed heart rate variability as a consistent biomarker for depression.

4.3 XAI in Prediction of Affective Disorders from EEG

There have already been some works investigating XAI in the prediction of affective disorders from EEG data, with most studies focusing on depression and high-performing deep-learning models. Below are briefly presented results of several relevant studies.

- One study [35] presented two deep model architectures that enable model visualization insights into spectral features learned by models and the spectral feature distribution across channels. They found higher β power, potentially higher δ power, and higher brain-wide correlation most strongly represented within the right hemisphere for individuals diagnosed with major depressive disorder.

- Another study [36], employed feature extraction from the frontal EEG sub-bands, with features including Higuchi's fractal dimension and sample entropy. They

tried multiple classifiers on these features, including decision tree, linear discriminant analysis, kNN, random forest, and XGBoost. The results showed increased complexity in depression compared with healthy controls, and its increase with the deepening of depression. They identified features with the highest impact are dominantly high-frequency features.

- Additional paper [37] extracted Lempel–Ziv complexity and frequency domain feature power spectral density features from EEG. They studied the effects of different brain regions and region combinations with eyes closed and opened in a resting state to determine the presence of depression. They found the temporal region on its own and frontal, temporal, and central regions combined show the best predictive power.

5 Methodology

This chapter introduces the main methods used in this thesis to model and explain the prediction of affective disorders, how they were used and the rationale for choosing them.

5.1 Experiment Design

As the main objective was to increase the theoretical understanding of depression, the methods were selected to emphasize interpretability, aiming to provide a broad, high-level comprehension rather than focusing on explainability, which offers specific example-based understanding. Interpretability was considered from a **feature ranking** perspective. Different methods were chosen for feature ranking in order to determine which features could be most telling in the prediction of depression. Feature ranking is a natural way to comprehend machine learning models, allowing experiments with emphasis on comparison and usability rather than relying on human subject testing for explainability. With this in mind, the next ranking methods were employed:

1. **Feature correlation with diagnosis**
2. **Decision tree feature importance**
3. **Shapley additive explanations (SHAP)**

The secondary objective of this thesis was to compare different XAI methods to identify their advantages and disadvantages. Thus each of the methods above was chosen to represent one of the stages in the machine learning pipeline as presented in 2.4.3: pre-model, intrinsic and post-hoc XAI.

In addition to these methods, **model training and testing on subsets of data** was conducted, and the performance of models on different feature subsets was used as a possible alternative indicator of feature importance.

5.2 Experimental Setup

An outline of the experimental procedures is provided below.

1. **Preliminary Analyses:** General data analysis was conducted to understand the dataset and its characteristics. Hypothesis testing was performed to choose appropriate methods for further analyses of this dataset.

2. **Correlation-based Feature Ranking:** The correlation between features and the depression diagnosis was analyzed and ranked.

3. **Model-based Feature Ranking:** A decision tree model and an SVM model were trained and tested using all features. For SVM, data was standardized before training and testing. Both models underwent hyperparameter optimization. Gini importance was extracted from the decision tree and SHAP values were extracted from the SVM model to rank features. Each model was evaluated with accuracy, precision, recall, and F1-score.

4. **Feature Subset Evaluation:** Features were subsetted in several different ways. First, domain-informed subsetting included considering features which only concern specific electrodes, brain waves, or feature extraction methods. For example, features were filtered to include only those recorded on a specific electrode. Additionally, subsetting based on naturally imposed strategies was conducted – based on the literature review, based on hypothesis testing results and based on feature rankings. For all subsets decision tree and SVM were trained without hyperparameter optimization. Each model was evaluated with accuracy, precision, recall, and F1-score to identify the most effective combination of features for depression diagnosis. This resulted in 80 model evaluations: 40 subsets x 2 model types. This included 19 electrode-, 5 brain wave type-, 6 feature extraction method-, 6 literature-, 1 hypothesis testing-, and 3 feature ranking-based feature subsets.

All experiments were conducted using *Python* (version 3.12), *Conda* (version 4.13), and *Jupyter Notebooks* (version 6.4). Key libraries included *sklearn* for machine learning, *matplotlib* and *seaborn* for data visualization, *scipy* for hypothesis testing, and *shap* for SHAP computation. A random seed was set to ensure reproducibility in all experiments.

5.3 Chosen Methods

5.3.1 Hypothesis Testing

Hypothesis testing is a statistical method used to make inferences about a population based on sample data. It involves formulating two competing hypotheses: the null hypothesis (H_0), which suggests no significant difference or relationship, and the alternative hypothesis (H_1), which posits the opposite. During preliminary analyses, testing was performed to determine the nature of the dataset and choose methods with fitting assumptions for further analyses.

Shapiro-Wilk test was used to assess the normality of the feature distributions. The null hypothesis of this test is that the population is normally distributed. This particular test was chosen for its sensitivity to deviations from normality in small to medium-sized samples.

Wilcoxon rank-sum test was used to compare the central tendency of features between healthy and depressed individuals because the dataset was determined to mostly not follow normal distribution, making Student's T test unsuitable. The null hypothesis of this test is the probability of X being greater than Y is equal to the probability of Y being greater than X for randomly selected values X and Y from two populations.

Levene's test was used to assess the equality of variances in features between healthy and depressed individuals. It tests the null hypothesis that the population variances are equal. This test is robust against departures from normality and is thus suitable for comparing variances across these groups.

5.3.2 Correlation Analysis

Correlation is one possible method of understanding and analyzing relationships within data. It is a bivariate measure of the strength and direction of a relationship between two variables. The value of correlation typically lies between -1 and 1, with 0 representing no correlation, and the closer the value to 1 in absolute terms, the stronger the correlation. The direction of the relationship is indicated by the sign where positive indicates high values of one variable associated with high values of the other variable, and negative indicates high values of one variable associated with low values of the other variable.

Pearson's correlation coefficient, given by 5.1, is the most common method used for measuring the linear relationship of two variables. In addition to the linear relationship, it assumes the normality of variables. It is computed with the following formula:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}} \tag{5.1}$$

Spearman's rank correlation is another commonly used method, which computes the correlation between the ranked variables instead of raw measurements, as shown in 5.2 [38]. It assumes a monotonic relationship but no assumptions on linearity or normality, making it more robust to outliers than Pearson's correlation coefficient.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{5.2}$$

where $d_i = R(X_i) - R(Y_i)$ is the difference in the ranks of corresponding variables of each observation and n is the number of pairs of observations.

Point-biserial correlation is a correlation coefficient used when one variable is dichotomous, making it suitable for calculation of correlation between continuous and categorical attributes [39]. Its calculation is represented in 5.3.

$$r_{pb} = \frac{\bar{X}_1 - \bar{X}_0}{s_{n-1}} \sqrt{\frac{n_1 n_0}{n(n-1)}} \tag{5.3}$$

where $\bar{X}_1$ represents the mean value of the continuous variable X for group 1, likewise for $\bar{X}_0$, and s_{n-1} represents the standard deviation: $s_{n-1} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$

Pearson's correlation was used in this thesis for reference during exploratory data analysis of relationships among features because it is the most used correlation measure. However, it is not fit for the given dataset because the normality assumption is not met. Thus Spearman's rank correlation was additionally used. For feature correlation with diagnosis, neither of these was fit, as they are designed for measuring the correlation of continuous variables, and this required a correlation measure between a continuous and a categorical variable. For this reason point-biserial correlation was used.

5.3.3 Decision Tree

Decision trees are the most popular interpretable algorithm for classification and regression [15, 21]. The name of these methods is justified by a tree-like structure, where inner nodes correspond to features, branches beneath nodes correspond to values of the features, and leaves correspond to classification decisions (class labels). A simple example of such a tree is represented in 5.1. An instance is classified by comparing its feature values against tree branches, starting from the root and traversing down to the leaves, until a leaf is reached and the corresponding label is assigned.

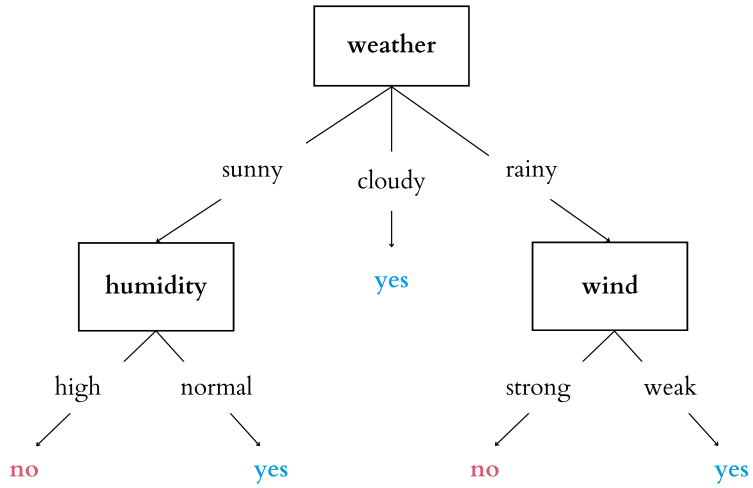


Figure 5.1: Decision tree for "a day for beach volleyball"

Building a decision tree involves recursively partitioning the dataset based on features that best separate the data into the purest subsets with respect to class labels. The process starts with the entire dataset. The algorithm evaluates all features and chooses one that best separates the data into two homogeneous groups using metrics like gini impurity or information gain. Once the feature for the split is determined, branches are created for each unique value of that feature. This splitting process is repeated for each subset until no further splitting is necessary or a predefined stopping criterion is met.

Gini impurity measures the likelihood of an incorrect classification of a randomly chosen element if it was randomly labeled according to the distribution of class labels in the

subset. A lower gini impurity indicates that the subset is more homogeneous, meaning that a larger proportion of the data points belong to the same class. The gini impurity is mathematically given by the formula 5.4.

$$Gini(D) = 1 - \sum_{i=1}^{k} p_i^2 \tag{5.4}$$

where k is the number of classes, and p_i is the proportion of instances in D that belong to class i.

Information gain measures the reduction in entropy achieved by partitioning the data based on a given attribute. Entropy quantifies the uncertainty or randomness in a dataset with respect to class labels, and the goal is to maximize the reduction in entropy when splitting the data. Mathematically, this is calculated as presented in 5.5.

$$IG(D, x) = E(D) - \sum_{V \in V(x)} \frac{|D_{x=v}|}{|D|} E(D_{x=v}) \tag{5.5}$$

where $E(D) = -\sum_{i=1}^{k} p_i log_2 p_i$ is the entropy of dataset D, $V(x)$ is the set of possible values for feature x, and D_P is the subset of examples satisfying condition P.

These measures are also used to compute the feature importance of a decision tree model. A feature's importance equals the total decrease in gini impurity or increase in information gain attributed to each feature over all the nodes where it is used to make decisions.

The simplicity of decision trees and their straightforward method for computing feature importance make them highly intuitive and easy to explain. For this reason, they were selected as the method of choice for the intrinsic explainability stage in the machine learning pipeline.

5.3.4 Support Vector Machine (SVM)

Support vector machines (SVMs) are a class of supervised learning algorithms well-known for their ability to handle complex decision boundaries in high-dimensional spaces.

SVMs seek to find the optimal hyperplane that best separates different classes of data

points. This hyperplane is positioned in such a way that it maximizes the margin between the nearest data points of different classes, known as support vectors [15, 16]. The support vectors define the decision boundary and provide memory efficiency by eliminating the need for large amounts of internal model parameters compared to other models.

Additional strength of SVMs lies in their ability to handle both linearly separable and non-linearly separable data. This is achieved through the use of the kernel trick, which allows SVMs to implicitly map the input data into higher-dimensional feature spaces where it might be easier to separate the classes linearly. However, this transformation makes them less interpretable, which is why they are often considered "black box" models. For this reason, this machine learning algorithm was chosen for the basis model on top of which the post-hoc XAI method was performed. Because it is treated as a "black box" within this thesis, the computational framework behind this method is omitted.

SVM was additionally a good choice for modeling within this thesis because it handles high-dimensional data well, and the dataset at hand contains a disproportionately large amount of features as opposed to the amount of examples. SVM also doesn't suffer from multicollinearity, which is another issue of this dataset further explained in the results section.

5.3.5 SHapley Additive exPlanations (SHAP)

Shapley values can be used as a post-hoc method for explaining the feature importance of a machine learning model [40, 41]. The concept behind these values is rooted in cooperative game theory, and the goal is to distribute the total gain among players based on their contribution to the overall outcome. For a given prediction, the Shapley value of a feature represents its average marginal contribution across all possible subsets of features. This involves computing the marginal contribution of a feature by adding it to a subset of features and measuring the change in the model's prediction. This process is repeated for all subsets, and the contributions are averaged to obtain the Shapley value.

To illustrate, consider a simple model with three features $\{A, B, C\}$. To compute the Shapley value for feature A, the impact of adding A to every possible subset of the other features needs to be evaluated, such as $\{B, C\}$, $\{B\}$, $\{C\}$. However, all permutations need to be considered: all orders in which A can be added to these subsets, such as (B, C, A),

(B, A, C), (A, B, C), etc. The Shapley value is then the weighted average of the changes in a model's prediction, considering all permutations and subsets.

The formula 5.6 illustrates the mathematical computation of the Shapley value $\phi_i(v)$ for feature i. Here, $v(S)$ represents the model's prediction based on a subset of features S, and N represents the set of all features. The equation sums over all subsets S that exclude feature i, weighting each subset's contribution by a factor that accounts for the number of permutations of feature sets. The term $v(S \cup \{i\}) - v(S)$ captures the change in the model's prediction when feature i is added to subset S. Thus, $\phi_i(v)$ provides a fair measure of player i's marginal contribution to the collective value v.

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S)) \tag{5.6}$$

Despite the obvious advantages of SHAP, the computation of Shapley values poses significant challenges due to computational complexity. The exact calculation of Shapley values requires evaluating all possible subsets of features, which is computationally infeasible for models with a large number of features. To mitigate this, several approximation methods have been developed, such as the Kernel SHAP and Tree SHAP algorithms.

5.4 Rejected Methods

Several methods were considered and tested for this thesis but rejected in the final experiment design:

Logistic Regression was considered as an intrinsically interpretable model. The weights of a fitted regression model can be used to imply feature importance under certain circumstances, mainly the absence of multicollinearity. As the dataset has a high amount of highly correlated features, it wasn't surprising when initial testing showed its low performance. Assessing explainability only makes sense on high-performing models, as they capture a pattern in the data well, and an explanation of the pattern is of interest. Since logistic regression did not meet the performance criteria, it was not pursued further for explainability analysis in this study.

Deep Models were considered as "black box" models in the post-hoc explainability test-

ing. However, they didn't show a large improvement in performance as opposed to the more simple models and were thus replaced by SVM models which also capture non-linear relationships well, but are faster and easier to employ.

Greedy Feature Selection Methods prioritize optimizing a performance metric, such as minimizing error or maximizing accuracy, during the selection process. They iteratively select features based on their individual contributions to the metric, often depending on the order in which features are evaluated or added. This sequential dependency can influence which features are ultimately chosen, impacting the overall robustness and generalizability of the selected feature subset.

LIME was considered for the model-agnostic XAI approach. However, it offers local explanations that were not the focus of this thesis, and thus it was not utilized.

Permutation Feature Importance was another model-agnostic method considered. It is global in character but preliminary testing showed it usually selected a small amount of features, it didn't give a ranking of many features, which made it less comparable to other methods chosen. SHAP is also considered an upgrade of permutation feature importance, so the two strategies would most likely show similar results.

5.5 Limitations

5.5.1 Formal XAI Evaluation

Due to time and resource constraints, a formal evaluation of XAI methods on users could not be conducted. Evaluating the user-centric effectiveness of XAI methods was also not the primary goal of this thesis; instead, the focus was on exploring different methodologies and their implications for understanding the nature of depression itself. Consequently, the study prioritized naturally understandable concept of feature importance ranking, which was compared across different methods to gain insights into EEG-based biomarkers of depression.

5.5.2 Imbalanced Dataset

The dataset used in this study exhibits significant class imbalance with a predominant representation of depressed individuals compared to healthy individuals. This imbal-

ance inherently introduces bias in model predictions towards the majority class. To address the challenges posed by the imbalanced dataset, the F1-score was selected as the primary metric for model evaluation.

5.5.3 High-Dimensional, Low-Sample Size Dataset

The dataset used in this study is characterized by a high number of features relative to the number of samples, fitting the profile of a High-Dimensional, Low-Sample Size (HDLSS) dataset. This presents several challenges for model training and evaluation, primarily due to the risk of overfitting and the curse of dimensionality.

5.5.4 Single Train-Test Split

The dataset utilized in this study was specifically designed for medical application, incorporating one train-test split that aims to represent the target population well in terms of age and sex matching. While this design choice ensures demographic consistency, it limits the ability to assess model variability and generalizability through techniques such as cross-validation.

Typically, multiple train-test splits or cross-validation would provide a more comprehensive assessment of model performance across different subsets of the data. The reliance on a single train-test split restricts the ability to generalize findings beyond the specific partition of data used in this study. Consequently, the strength of claims and conclusions within this thesis may be limited due to the lack of variability assessment across different data partitions.

5.5.5 Multiple Comparisons Problem

In the process of evaluating numerous features and their combinations, the study inherently faced the multiple comparisons problem. Training a large amount of models naturally leads to a greater risk of finding at least one good-performing one purely by chance. This would be easier to mitigate with more data and multiple train-test splits.

6 Results

6.1 Preliminary Analyses

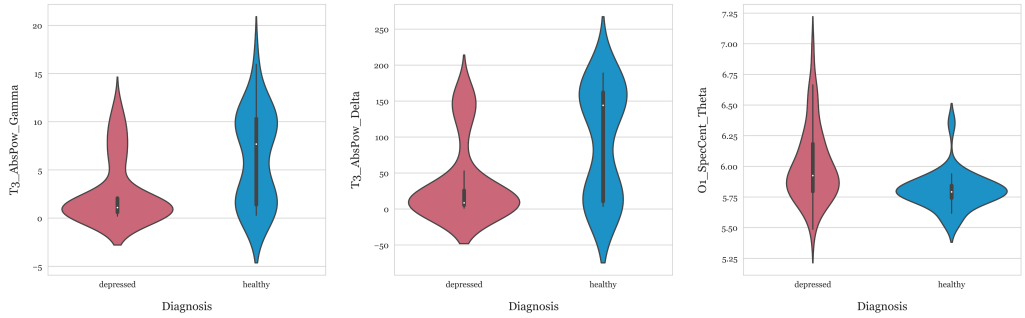
Traditional visualization-focused exploratory data analysis was not possible due to the high amount of features in the dataset. Consequently, various visualization techniques were applied selectively to individual features or subsets of the data. Violin plots were used to examine features which hypothesis testing showed to be interesting in some way. Multiple correlation heatmaps were employed in an attempt to unveil relationships within the data.

6.1.1 Hypothesis Testing

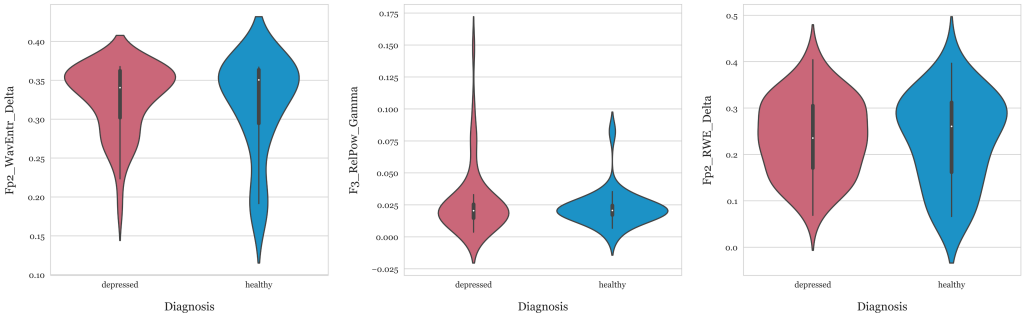
All conclusions were based on a significance level of **0.05**:

- Normality testing showed none of the features follow a normal distribution.
- Central tendency testing demonstrated a statistically significant difference in central tendency between healthy and depressed individuals across 159 out of 570 features. List of these features can be found in appendix A.
- Homogeneity of variances testing revealed significant differences in variance across all features between healthy and depressed individuals.

The distributions of the three most significantly and three least significantly different features in terms of central tendency are shown in figure 6.1. Interestingly, many features showed bimodal distribution for healthy individuals and a distinctly more unimodal distribution of the same features for depressed individuals.



(a) Most significantly different



(b) Most insignificantly different

Figure 6.1: Distributions of selected features according to central tendency difference

6.1.2 Correlation Among Features

Correlation analysis showed high multicollinearity within the dataset. Out of a total of 162165 possible pairs of features in the dataset, there are many very highly correlated pairs (correlation larger than **0.99**):

- **377** according to Pearson,
- **92** according to Spearman.

To further illustrate the extent of this high amount of correlated features, an examination was conducted on features correlated with a large number of other features. The number of features correlated with at least **10 other features** with a minimum of **0.9** correlation was observed as follows:

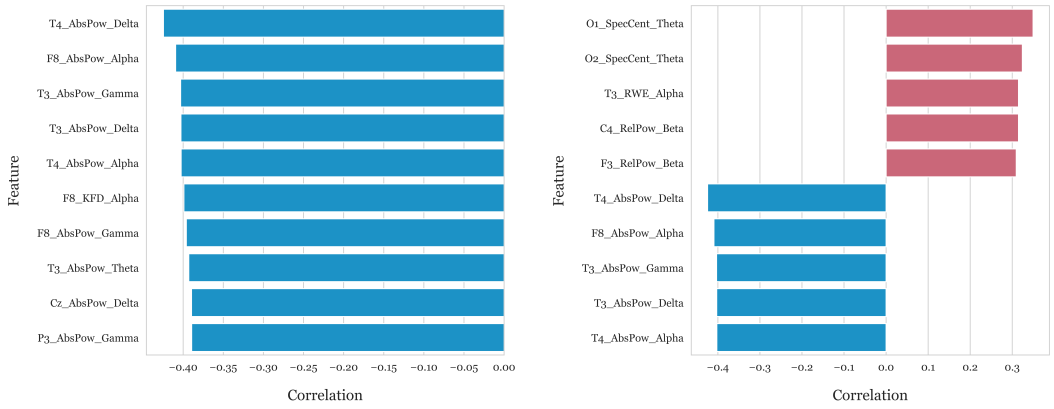
- **163** according to Pearson,
- **124** according to Spearman.

6.2 Correlation-Based Feature Ranking

Generally, according to Point-biserial correlation measure, there are more features negatively correlated with the diagnosis than there are positively correlated features. The ten most correlated features in the absolute sense are all negatively correlated. The highest correlation rankings are presented in figure 6.2.

Most of the highly correlated features are from the temporal electrodes T3 and T4, with three additional features from the frontal electrode F8. Additionally, there is one feature each from the central electrode Cz and the parietal electrode P3. In terms of brain waves, delta, alpha, and gamma waves are equally represented, each contributing three features, while one feature is derived from the theta wave. The beta wave is not represented. Notably, nine out of the ten features are related to absolute power, with the tenth feature associated with Katz fractal dimension.

Among the positively correlated features, notable are the spectral centroids from theta waves on occipital electrodes.



(a) Top 10 in absolute sense

(b) Top 5 positively and top 5 negatively correlated features

Figure 6.2: Most correlated features with the diagnosis

6.3 Model-Based Feature Ranking

Both decision tree and SVM models outperform the majority classifier in terms of accuracy, as shown in 6.1. Decision tree performs better with an accuracy of 73% and F1-score of 82% as compared to SVM with an accuracy of 72% and F1-score of 80%.

Table 6.1: Model performances on all features

Classifier	**Accuracy**	**Precision**	**Recall**	**F1-score**
Majority Classifier	0.67	0.67	1.00	0.80
Decision Tree	0.73	0.75	0.90	**0.82**
Support Vector Machine	0.70	0.72	0.90	0.80

Decision trees don't provide feature ranking for all features, they only rank features present in the constructed tree. For this dataset the constructed tree is represented in 6.3. Non-present features can be considered to have a zero importance.

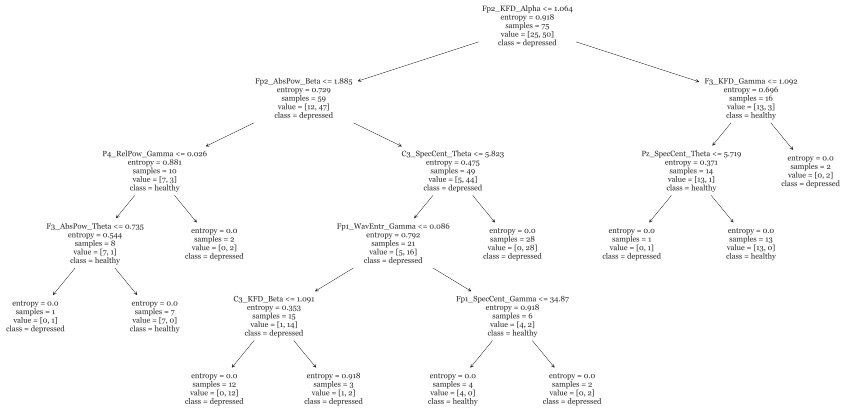


Figure 6.3: Decision tree for depression diagnosis illustrated

Decision tree feature ranking is provided in figure 6.5(a). The two most important features originate from the same prefrontal electrode, Fp2. Additionally, two features come from another prefrontal electrode, Fp1, two from the frontal electrode F3, and two from the central electrode C3. Notably, no features from temporal, parietal, or occipital electrodes are present. Regarding brain waves, the alpha wave contributes to the most important feature. The gamma wave is the most frequently represented, with four features, followed by theta with three features, and beta with two features. The delta wave is not represented at all. Among the feature extraction methods, Katz fractal dimension stands out with three features, matched by spectral centroids, while other methods are less prominent, with relative wavelet energy entirely absent.

SHAP values based on the trained SVM model are presented in its original visualization in figure 6.4, and adjusted visualization for comparability in figure 6.5(b). Each dot on

the original plot corresponds to a SHAP value for a specific feature in a single observation, and the dots are color-coded based on the feature's value, red for high and blue for low values. This color gradient helps in understanding the relationship in terms of both magnitude and direction between feature values and their impact on the model's output. A wider spread of dots along the x-axis shows more variability in the feature's influence across different predictions. Interestingly, the most important feature, absolute power of delta on O1 electrode, is also the one with the widest spread, with extreme cases for depressed individuals.

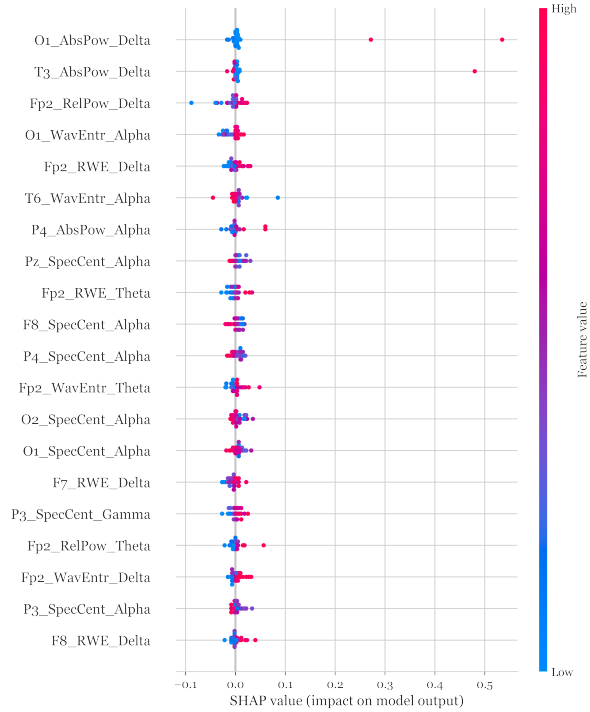
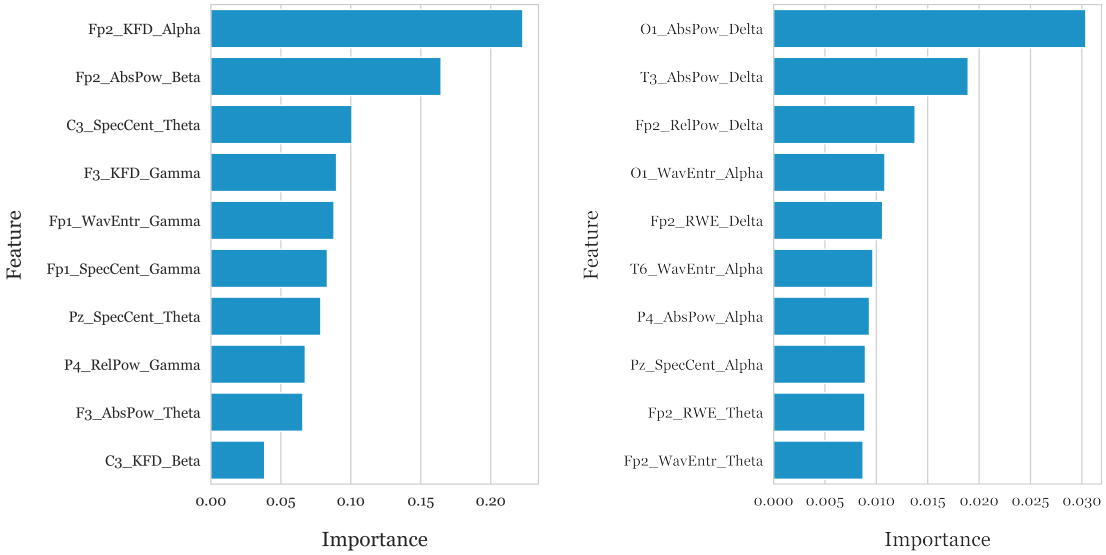


Figure 6.4: Effect of features on SVM model performance – SHAP values

SHAP values are non-zero for every feature, so to facilitate comparability, a cut-off was applied to present only the top 10 most important features. In comparability adjusted plot, it is visible that two highly important features are based on the occipital electrode O1. The prefrontal electrode Fp2 is also prominently represented. Regarding the types of waves, four features concern delta waves, with three of them ranking at the top in importance. Additionally, there are four features related to alpha waves and two to theta waves, while beta and gamma waves are not represented. In terms of feature types, three are based on absolute power, three on wavelet entropy, two on relative wavelet energy, one on relative power, and one on spectral centroids. The Katz fractal dimension is no-

tably absent among the represented features.



(a) Decision Tree

(b) SHAP

Figure 6.5: Feature rankings based on models

6.4 Feature Subset Evaluation

6.4.1 Domain-Informed Subsetting

In contrast with commonly used datasets where a comprehensible relationship between features is unknown or non-existent, this EEG-derived dataset holds feature names according to different feature extraction methods, brain waves they were extracted from, and the electrode they were recorded on. Using this knowledge, subsets of features were constructed to discern how they affect model performance, giving additional insight and indication of where the predictive power within the dataset comes from. The analysis is categorized into three groups: per feature extraction method, per brain wave type, and per electrode placement.

By Feature Extraction Method

Results by feature extraction method are shown in 6.2. For absolute band power, decision tree outperforms SVM, whereas for other methods SVM tends to slightly outperform. Spectral centroid, Relative wavelet energy and wavelet entropy subsets indicate larger predictive capability for both models.

By Brain Wave

Results by brain wave type are shown in 6.3. Subsetting by brain wave seems to generally negatively effect the performance of the decision tree, indicating relevance of interactions among features extracted from different brain waves. With exception of beta wave which seems to hold important information for decision tree, resulting in a high F1-score of 88%. SVM seems to handle this lack better, learning the patterns from only one brain wave subset well. It performed better on delta, theta, and beta subsets.

By Electrode

Results by electrode are shown in 6.4. Performance across electrodes and models varies. Electrodes which show high predictive power for both models are Fp1, F7, C4, P3 and P4. Electrode displaying least predictive power is T5. In general, decision tree seems to suffer from a lack of other electrodes more than SVM, which maintains a good performance for most subsets, but never achieves as high F1-score as the decision tree in some cases.

Table 6.2: Model performances on feature subsets by feature extraction method

Feature Subset	Classifier	Accuracy	Precision	Recall	F1-score
	Majority Classifier	0.67	0.67	1.00	0.80
Absolute Band Power	Decision Tree	0.77	0.78	0.90	0.84
	Support Vector Machine	0.63	0.67	0.90	0.77
Relative Band Power	Decision Tree	0.60	0.68	0.75	0.71
	Support Vector Machine	0.73	0.73	0.95	0.83
Spectral Centroid	Decision Tree	0.77	0.78	0.90	0.84
	Support Vector Machine	0.77	0.76	0.95	0.84
Relative Wavelet Energy	Decision Tree	0.73	0.73	0.95	0.83
	Support Vector Machine	0.77	0.74	1.00	**0.85**
Wavelet Entropy	Decision Tree	0.73	0.73	0.95	0.83
	Support Vector Machine	0.77	0.74	1.00	**0.85**
Katz Fractal Dimension	Decision Tree	0.63	0.71	0.75	0.73
	Support Vector Machine	0.73	0.77	0.85	0.81

Table 6.3: Model performances on feature subsets by brain wave type

Feature Subset	Classifier	Accuracy	Precision	Recall	F1-score
	Majority Classifier	0.67	0.67	1.00	0.80
Delta	Decision Tree	0.43	0.60	0.45	0.51
	Support Vector Machine	0.73	0.73	0.95	0.83
Theta	Decision Tree	0.47	0.60	0.60	0.60
	Support Vector Machine	0.73	0.75	0.9	0.82
Alpha	Decision Tree	0.67	0.69	0.90	0.78
	Support Vector Machine	0.70	0.74	0.85	0.79
Beta	Decision Tree	0.83	0.83	0.95	**0.88**
	Support Vector Machine	0.73	0.75	0.9	0.82
Gamma	Decision Tree	0.70	0.82	0.70	0.76
	Support Vector Machine	0.70	0.74	0.85	0.79

Table 6.4: Model performances on feature subsets by electrode

Feature Subset	Classifier	Accuracy	Precision	Recall	F1-score
	Majority Classifier	0.67	0.67	1.00	0.80
Fp1	Decision Tree	0.73	0.75	0.90	0.82
	Support Vector Machine	0.77	0.76	0.95	0.84
Fp2	Decision Tree	0.63	0.68	0.85	0.76
	Support Vector Machine	0.70	0.72	0.90	0.80
F7	Decision Tree	0.83	0.83	0.95	**0.88**
	Support Vector Machine	0.73	0.75	0.90	0.82
F3	Decision Tree	0.57	0.67	0.70	0.68
	Support Vector Machine	0.73	0.77	0.85	0.81
Fz	Decision Tree	0.63	0.70	0.80	0.74
	Support Vector Machine	0.73	0.77	0.85	0.81
F4	Decision Tree	0.73	0.77	0.85	0.81
	Support Vector Machine	0.73	0.77	0.85	0.81
F8	Decision Tree	0.60	0.67	0.80	0.73
	Support Vector Machine	0.77	0.78	0.90	0.84
T3	Decision Tree	0.63	0.70	0.80	0.74
	Support Vector Machine	0.70	0.74	0.85	0.79
C3	Decision Tree	0.67	0.73	0.80	0.76
	Support Vector Machine	0.77	0.76	0.95	0.84
Cz	Decision Tree	0.57	0.65	0.75	0.70
	Support Vector Machine	0.73	0.75	0.90	0.82
C4	Decision Tree	0.77	0.81	0.85	0.83
	Support Vector Machine	0.73	0.75	0.90	0.82
T4	Decision Tree	0.63	0.71	0.75	0.73
	Support Vector Machine	0.77	0.78	0.90	0.84
T5	Decision Tree	0.53	0.67	0.60	0.63
	Support Vector Machine	0.67	0.71	0.85	0.77
P3	Decision Tree	0.83	0.86	0.90	**0.88**
	Support Vector Machine	0.77	0.78	0.90	0.84
Pz	Decision Tree	0.67	0.75	0.75	0.75
	Support Vector Machine	0.77	0.76	0.95	0.84
P4	Decision Tree	0.73	0.80	0.80	0.80
	Support Vector Machine	0.73	0.75	0.90	0.82
T6	Decision Tree	0.63	0.76	0.65	0.70
	Support Vector Machine	0.73	0.75	0.90	0.82
O1	Decision Tree	0.77	0.78	0.90	0.84
	Support Vector Machine	0.70	0.74	0.85	0.79
O2	Decision Tree	0.57	0.66	0.75	0.70
	Support Vector Machine	0.70	0.72	0.90	0.80

6.4.2 Additional Subsetting

Over the course of other analyses within the thesis, several additional subsetting strategies emerged as ideas organically.

Firstly, inspired by findings from previous literature, as detailed in 4.3, which suggested differences in brain activity patterns between healthy and depressed individuals in frontal regions, significant effects in the right hemisphere, and higher β and γ power in depressed individuals. Naturally, the question whether this holds true for this dataset posed itself. While some of these aspects were addressed by previous subsetting strategies, the specific exploration of region-based information had not been explored.

After hypothesis testing showed significant differences in central tendency between healthy and depressed individuals for a fraction of features, it naturally prompted the idea of evaluating models using only these discriminatory features.

Lastly, after ranking features based on criteria such as correlation, decision tree importance, or SHAP values, there arose the opportunity to evaluate models using only the top-ranked features. This strategy aimed to explore whether prioritizing features deemed important by these ranking methods could lead to more effective models in distinguishing between healthy and depressed individuals.

a) Literature Informed

Results on this dataset didn't entirely show agreement with previous literature. Frontal electrodes seem to hold predictive power for SVM but not for decision trees. A subset from prefrontal and frontal electrodes together performed the same on both models, implying interactions within frontal and prefrontal regions are not important for the prediction of depression. Even more diverging results showed right hempisphere based electrodes didn't hold more predictive power. On the contrary, the left hemisphere showed better performance and when combined with central electrodes, it showed the best performance on the decision tree model over all experiments.

b) Hypothesis Testing Informed

The decision tree model performed quite well on hypothesis testing based preselected features.

c) Ranking Informed

Perhaps the most surprising results showed that ranking according to either correlation with diagnosis or high SHAP values is not very helpful as a feature selection method. On the other hand, the decision tree deemed important features showed good performance.

Table 6.5: Model performances on feature subsets based on a) literature research, b) hypothesis testing, c) highest ranked features

Feature Subset	**Classifier**	**Accuracy**	**Precision**	**Recall**	**F1-score**
	Majority Classifier	0.67	0.67	1.00	0.80
a) Frontal Electrodes	Decision Tree	0.70	0.76	0.80	0.78
	Support Vector Machine	0.77	0.78	0.90	0.84
Frontal & Prefrontal Electrodes	Decision Tree	0.70	0.76	0.80	0.78
	Support Vector Machine	0.77	0.78	0.90	0.84
Even Electrodes (right hemisphere)	Decision Tree	0.60	0.67	0.80	0.73
	Support Vector Machine	0.70	0.74	0.85	0.79
Even & Midline Electrodes	Decision Tree	0.63	0.70	0.80	0.74
	Support Vector Machine	0.70	0.73	0.85	0.79
Odd Electrodes (left hemisphere)	Decision Tree	0.80	0.89	0.8	0.84
	Support Vector Machine	0.70	0.72	0.90	0.80
Odd & Midline Electrodes	Decision Tree	0.87	0.87	0.95	**0.90**
	Support Vector Machine	0.70	0.72	0.90	0.80
b) Hypothesis Testing Informed	Decision Tree	0.83	0.83	0.95	**0.88**
	Support Vector Machine	0.70	0.72	0.90	0.80
c) High Correlation with Diagnosis	Decision Tree	0.63	0.74	0.70	0.72
	Support Vector Machine	0.63	0.68	0.85	0.76
High Decision Tree Importance	Decision Tree	0.77	0.76	0.95	**0.84**
	Support Vector Machine	0.73	0.75	0.90	0.82
High SHAP Value	Decision Tree	0.5	0.65	0.55	0.59
	Support Vector Machine	0.67	0.68	0.95	0.79

7 Discussion

This section aims to understand, hypothesize, and explain the underlying reasons for achieved results using employed methods. Additionally, it critically evaluates the benefits and weaknesses of these methods in the context of this thesis.

7.1 Preliminary Analyses

The initial analyses conducted on the dataset have unveiled several critical aspects that shape understanding of the data's characteristics and its implications. The dataset exhibited high dimensionality, multicollinearity among variables, and imbalance between healthy and depressed individuals.

Notably, significant differences in central tendency were observed across 159 features between healthy and depressed individuals, highlighting substantial variability that may contribute to predictive modeling efforts.

7.2 Feature Rankings

Results of feature rankings showed different models found different features important. Only the T3_AbsPow_Delta feature is shared, with high correlation and high SHAP value. In terms of brain hemispheres, correlation and Shapley prioritized the right hemisphere while the decision tree prioritized the left hemisphere. Interestingly, all three methods had one midline feature present. When aggregated among all three methods, important regions seem to be prefrontal features, followed by temporal, and then frontal, and parietal regions. In terms of brain waves, when aggregated, delta, alpha, and gamma seem to be most important.

Correlation provided a good initial understanding of relationships between features and

diagnosis but this approach does not account for interactions between features. It also assumes a monotonic relationship, which may not be true for the complex neural system that is our brain. For this reason, it is not sufficient to capture the full complexity of the data. This was further proven by initial testing of the logistic regression model mentioned in 5.4, indicating the relationship between the features and the diagnosis is most likely dominantly not linear in nature. Consequently, models built on subsets of features selected based on high correlation did not yield good performance. Regions most highly correlated with diagnosis are temporal region features, followed by frontal region features. The feature extraction method which showed a high negative correlation with diagnosis was absolute power. This implies that depressed individuals tend to have more pronounced brain waves in an absolute sense than healthy individuals.

Decision trees, on the other hand, recursively split the data based on features that best differentiate between diagnoses. This way interactions between features are implicitly captured, leading to a different set of important features compared to correlation analysis. Interestingly, the decision tree favours Katz fractal dimension and spectral centroids of left hemisphere electrodes, two measures that essentially measure irregularity, implying irregularity of the left hemisphere is relevant in the differentiation between healthy and depressed people. Evaluation of models on the subset of features most important to the decision tree showed good performance on both models.

SVMs aimed to find a hyperplane that best separates classes in a high-dimensional space defined by the features. Thus its feature rankings are determined by their contribution to defining the decision boundary rather than by their relationships with the diagnosis. This exhibits a different paradigm from correlation and decision tree and thus is not surprising that the rankings are different. Evaluation of models on the subset of features most important to SVM showed bad performance on both models.

7.3 Subset Model Performance

In terms of performance, the decision tree seems to have better performance the more different features are available, whereas SVM seems to be more robust in terms of performing well on various subsets of features.

7.3.1 Based on Feature Extraction Method

Spectral centroid, relative wavelet energy and wavelet entropy extraction method subsets showed good performance in both decision tree and SVM models. This implies that these extraction methods might effectively capture the information that differs between depressed and healthy individuals. For decision trees, they might offer a straightforward split criterion, and for SVMs, they might help define margins between healthy and depressed individuals.

Relative band power and Katz fractal dimension subsets showed good performance only on SVMs. For relative band power, this could be because these features exhibit gradual changes across different classes, which are better handled by SVMs that create smooth decision boundaries. Decision trees, which work by making binary splits, might struggle to capture these subtle variations effectively. For Katz fractal dimension, the complexity of the signal might be captured in a way that might lead to non-linear relationships between the feature values and the target classes. SVMs, especially with non-linear kernels, excel at modeling such relationships, while decision trees might oversimplify.

7.3.2 Based on Brain Wave Type

Subsetting features by brain wave types yielded intriguing results. It appears that decision trees benefit from a variety of different brain wave types. This could suggest that interaction among features extracted from different brain wave types could be important in understanding differences between depressed and healthy individuals. The exception from this are beta waves where decision tree performs well despite the lack of other wave types. Beta waves might contain more discriminative information than other brain wave types.

On the other hand, SVMs demonstrate a consistent performance across different brain wave subsets. This makes sense as they are designed to maximize the margin between classes, which generally leads to better generalization.

7.3.3 Based on Electrode

Evalution on electrode level subsets showed mainly prefrontal, frontal, central and parietal electrodes hold more predictive power than others. In general, decision tree seems to suffer from lack of other electrodes more than SVM, just like in the case of brain wave subsets. SVM maintains a good performance for most subsets, but never achieves as high F1-score as the decision tree in some cases.

7.4 Comparing Different XAI Methods

To address the second objective of this thesis, an examination of strengths, weaknesses and applications of different XAI methods is summarized.

7.4.1 Pre-Model: Correlation with Diagnosis

Correlation was chosen as the pre-model explainability method due to its simplicity in calculation and interpretation. It facilitates quick identification of potentially important features and the directionality of influence. However, its scope is limited as it does not account for interactions between features. Additionally, correlation does not imply causation; a high correlation does not necessarily indicate that a feature causes changes in diagnosis.

7.4.2 Intrinsic: Decision Tree

Decision trees were considered as intrinsically explainable method. Its benefits include an intuitive hierarchical structure and the ability to handle non-linearity. Its main downside seems to be instability, as small changes in subset of features can lead to large variations in performance and importance rankings. They also doesn't provide directional information.

7.4.3 Post-hoc: SHAP

SHAP was chosen as the post-hoc XAI method. Its main advantage lies in its model-agnostic nature, making it applicable across different types of models. SHAP provides both global and local interpretability by constructing explanations based on individual predictions which also allow aggregation. However, calculating SHAP values can be

computationally expensive, especially for large datasets or complex models. Moreover, its complexity relative to other methods may pose challenges for non-experts.

7.4.4 Bottomline

All things considered, all three methods show benefits for various scenarios. Correlation can be used when there is a need for a quick, initial assessment of feature importance. Decision tree importance should be used when hierarchical understanding of features is needed. SHAP values can be utilized when dealing with black-box methods, especially with deep models where other interpretative methods may not suffice. For scenarios where directional information is crucial, either correlation or SHAP values are recommended. Decision trees are of more use when rule-based information is needed.

However, in the case of understanding differences between depressed and healthy individuals, none of these methods seem entirely appropriate due to their reliance on assumptions that may not hold true.

8 Conclusion

This thesis employed an extensive analysis of EEG-derived features to explore their utility in distinguishing between healthy individuals and those diagnosed with depression. The investigation spanned various methodological approaches, including exploratory data analysis, hypothesis testing, correlation analysis, feature ranking, and feature subset evaluations. These efforts aimed to uncover meaningful patterns and interactions within the dataset and assess their predictive power in depression diagnosis. Additionally, different methods of explainability were employed to evaluate their agreement and identify important biomarkers of depression.

The findings suggest that explainability in machine learning models might not be the most effective approach for distinguishing between healthy and depressed individuals. Different XAI methods yielded inconsistent results due to their distinct assumptions and underlying model architectures, which restricted their usefulness in identifying general differences between healthy and depressed individuals.

Other performed analyses suffered from the size and feature-to-instance ratio of the dataset. With over five times the number of features compared to examples and a single representative train-test split, it was challenging to draw statistically significant conclusions. Despite these limitations, the analysis identified 159 features with significant differences in central tendencies and significant variance differences across all features between the two groups. No statistically significant conclusions were drawn from other analyses, although they can inform future research. Notably, the results suggest that the relationship between EEG features and depression diagnosis is non-linear in nature. The prefrontal, frontal, temporal and parietal brain regions together seem to hold higher importance than other regions. When used in isolation frontal region or left hemisphere combined with the midline may contain more predictive information than other regions,

and beta waves might also be particularly informative when investigated without other brain wave types. A combination of alpha, delta and gamma seem to be promising when used together.

The implications of this research include help in the future identification of EEG biomarkers that differentiate between healthy and depressed individuals and advances in understanding the neurophysiological underpinnings of depression. This knowledge could lead to more effective diagnostic tools, enabling earlier and more accurate identification of depression. Additionally, pinpointing specific brain regions or EEG frequency bands associated with depression could inform targeted therapeutic strategies, such as brain stimulation therapies tailored to modulate activity in these regions and frequency bands, potentially increasing their effectiveness. In the broader context of mental health research, this thesis highlights the importance of large, well-balanced datasets and robust analytical techniques for developing reliable diagnostic and prognostic tools. It also underscores the potential of machine learning and advanced statistical methods in uncovering complex, non-linear relationships within neurophysiological data.

Overall, although the dataset limitations prevented definitive conclusions, the insights provide a valuable foundation for future research. Studies with larger datasets and more robust statistical methods could deepen our understanding of the EEG characteristics linked to depression, enhancing predictive model development and ultimately improving mental health care and patient outcomes.

References

[1] *International Statistical Classification of Diseases and Related Health Problems*, 10th ed. World Health Organization, 1992.

[2] K. M. Smith, P. F. Renshaw, and J. Bilello, "The Diagnosis of Depression: Current and Emerging Methods," *Comprehensive Psychiatry*, vol. 54, no. 1, pp. 1–6, 2013.

[3] M. Hamilton, "A Rating Scale for Depression," *Journal of Neurology, Neurosurgery, and Psychiatry*, vol. 23, pp. 56–62, 1960.

[4] *International Statistical Classification of Diseases and Related Health Problems*, 11th ed. World Health Organization, 2021.

[5] Encyclopædia Britannica, "electroencephalography." https://www.britannica.com/science/electroencephalography, Accessed April 12th, 2024.

[6] F. H. Duffy, V. G. Iyer, and W. W. Surwillo, *Clinical Electroencephalography and Topographic Brain Mapping: Technology and Practice*. Springer Science & Business Media, 1989.

[7] S. Hitziger, "Modeling the variability of electrical activity in the brain," Ph.D. dissertation, Université Nice Sophia Antipolis, 2015.

[8] S. Nagel, "Towards a home-use BCI: fast asynchronous control and robust non-control state detection," Ph.D. dissertation, Universität Tübingen, 2019.

[9] V. Jurcak, D. Tsuzuki, and I. Dan, "10/20, 10/10, and 10/5 Systems Revisited: Their Validity as Relative Head-Surface-Based Positioning Systems," *Neuroimage*, vol. 34, no. 4, pp. 1600–1611, 2007.

[10] N. N. Boutros, S. Galderisi, O. Pogarell, and S. Riggio, *Standard Electroencephalography in Clinical Psychiatry: A Practical Handbook*. John Wiley & Sons, 2011.

[11] S. Sanei and J. A. Chambers, *EEG Signal Processing and Machine Learning*. John Wiley & Sons, 2021.

[12] C.-H. Im, *Computational EEG Analysis: Methods and Applications*. Springer Singapore, 2018.

[13] L. Hu and Z. Zhang, *EEG Signal Processing and Feature Extraction*. Springer Singapore, 2019.

[14] A. K. Singh and S. Krishnan, "Trends in EEG signal feature extraction applications," *Frontiers in Artificial Intelligence*, vol. 5, p. 1072801, 2023.

[15] E. Alpaydin, *Introduction to Machine Learning*. The MIT Press, 2010.

[16] J. Šnajder and B. D. Bašić, "Strojno učenje," 2014, unpublished manuscript, Faculty of Electrical Engineering and Computing, University of Zagreb.

[17] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information fusion*, vol. 58, pp. 82–115, 2020.

[18] G. Vilone and L. Longo, "Explainable Artificial Intelligence: a Systematic Review," *arXiv preprint arXiv:2006.00093*, 2020.

[19] Z. C. Lipton, "The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[20] "Software and systems engineering, Software testing, Part 11: Guidelines on the testing of AI-based systems," International Organization for Standardization, Geneva, Switzerland, ISO, 2020.

[21] U. Kamath and J. Liu, *Explainable Artificial Intelligence: an Introduction to Interpretable Machine Learning*. Springer, 2021, vol. 2.

[22] Z. Chen, F. Xiao, F. Guo, and J. Yan, "Interpretable Machine Learning for Building Energy Management: A State-of-the-Art Review," *Advances in Applied Energy*, vol. 9, p. 100123, 2023.

[23] "Interpretability," https://www.mathworks.com/discovery/interpretability.html, Accessed June 15th, 2024.

[24] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," *arXiv preprint arXiv:1702.08608*, 2017.

[25] E. Jovičić, A. Jović, and M. Cifrek, "Impact of EEG Signal Preprocessing Methods on Machine Learning Models for Affective Disorders," 2024.

[26] I. Kinder, K. Friganović, J. Vukojević, D. Mulc, T. Slukan, D. Vidović, P. Brečić, and M. Cifrek, "Comparison of Machine Learning Methods in Classification of Affective Disorders," in *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*. IEEE, 2020, pp. 177–181.

[27] J. Vukojević, D. Mulc, I. Kinder, E. Jovičić, K. Friganović, A. Savić, M. Cifrek, and D. Vidović, "Borderline and Depression: A Thin EEG Line," *Clinical EEG and neuroscience*, vol. 54, no. 3, pp. 224–227, 2023.

[28] M. F. Pinto, A. Leal, F. Lopes, J. Pais, A. Dourado, F. Sales, P. Martins, and C. A. Teixeira, "On the Clinical Acceptance of Black-Box Systems for EEG Seizure Prediction," *Epilepsia Open*, vol. 7, no. 2, pp. 247–259, 2022.

[29] M. Kiani, J. Andreu-Perez, H. Hagras, S. Rigato, and M. L. Filippetti, "Towards Understanding Human Functional Brain Development with Explainable Artificial Intelligence: Challenges and Perspectives," *IEEE Computational Intelligence Magazine*, vol. 17, no. 1, pp. 16–33, 2022.

[30] A. Sujatha Ravindran and J. Contreras-Vidal, "An Empirical Comparison of Deep Learning Explainability Approaches for EEG Using Simulated Ground Truth," *Scientific Reports*, vol. 13, no. 1, p. 17709, 2023.

[31] M. Z. Uddin, K. K. Dysthe, A. Følstad, and P. B. Brandtzaeg, "Deep Learning for Prediction of Depressive Symptoms in a Large Textual Dataset," *Neural Computing*

and Applications, vol. 34, no. 1, pp. 721–744, 2022.

[32] E. Kerz, S. Zanwar, Y. Qiao, and D. Wiechmann, "Toward Explainable AI (XAI) for Mental Health Detection Based on Language Behavior," *Frontiers in psychiatry*, vol. 14, p. 1219479, 2023.

[33] A. Alotaibi and C. Sas, "Review of AI-Based Mental Health Apps," in *36th International BCS Human-Computer Interaction Conference*. BCS Learning & Development, 2023, pp. 238–250.

[34] A. Faiz, "MoodAI: A Novel Explainable AI Framework for Depression," Master's thesis, University of Auckland, August 2023.

[35] C. A. Ellis, M. L. Sancho, R. L. Miller, and V. D. Calhoun, "Identifying EEG Biomarkers of Depression with Novel Explainable Deep Learning Architectures," *bioRxiv*, 2024.

[36] F. Chen, L. Zhao, L. Yang, J. Li, and C. Liu, "An Explainable Assessment for Depression Detection Using Frontal EEG," in *Asian-Pacific Conference on Medical and Biological Engineering*. Springer, 2023, pp. 377–383.

[37] J. Yang, Z. Zhang, P. Xiong, and X. Liu, "Depression Detection Based on Analysis of EEG Signals in Multi Brain Regions," *Journal of Integrative Neuroscience*, vol. 22, no. 4, p. 93, 2023.

[38] K. Ali Abd Al-Hameed, "Spearman's Correlation Coefficient in Statistical Analysis," *International Journal of Nonlinear Analysis and Applications*, vol. 13, no. 1, pp. 3249–3255, 2022.

[39] D. Kornbrot, "Point Biserial Correlation," *Wiley StatsRef: Statistics Reference Online*, 2014.

[40] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[41] C. Molnar, *Interpretable Machine Learning*. Leanpub, 2020.

A Hypothesis Testing Subset

Below listed are features which showed statistically significant difference in central tendencies between healthy and depressed individuals.

Fp1_WavEntr_Delta

Fp1_RelPow_Theta

Fp1_RWE_Theta

Fp1_WavEntr_Theta

Fp1_KFD_Alpha

Fp2_SpecCent_Beta

F7_AbsPow_Delta

F7_KFD_Delta

F7_AbsPow_Theta

F7_KFD_Theta

F7_AbsPow_Alpha

F7_KFD_Alpha

F7_AbsPow_Beta

F7_KFD_Beta

F7_AbsPow_Gamma

F7_KFD_Gamma

F3_AbsPow_Delta

F3_RelPow_Delta

F3_KFD_Delta

F3_AbsPow_Theta

F3_KFD_Theta

F3_AbsPow_Alpha

F3_KFD_Alpha

F3_RelPow_Beta

F3_RWE_Beta

F3_WavEntr_Beta

F3_KFD_Beta

F3_AbsPow_Gamma

F3_KFD_Gamma

Fz_AbsPow_Delta

Fz_RelPow_Delta

Fz_RWE_Delta

Fz_WavEntr_Delta

Fz_KFD_Delta

Fz_AbsPow_Theta

Fz_SpecCent_Theta

Fz_KFD_Theta

Fz_WavEntr_Alpha

Fz_KFD_Alpha

Fz_RelPow_Beta

Fz_AbsPow_Gamma

Fz_KFD_Gamma

F4_AbsPow_Delta

F4_KFD_Delta

F4_AbsPow_Theta

F4_SpecCent_Theta

F4_KFD_Theta

F4_AbsPow_Alpha

F4_KFD_Alpha

F4_RelPow_Beta

F4_RWE_Beta

F4_WavEntr_Beta

F4_AbsPow_Gamma

F4_KFD_Gamma

F8_AbsPow_Delta

F8_KFD_Delta

F8_AbsPow_Theta

F8_KFD_Theta

F8_AbsPow_Alpha

F8_KFD_Alpha

F8_AbsPow_Beta

F8_KFD_Beta

F8_AbsPow_Gamma

F8_KFD_Gamma

T3_AbsPow_Delta

T3_KFD_Delta

T3_AbsPow_Theta

T3_KFD_Theta

T3_AbsPow_Alpha

T3_RelPow_Alpha

T3_RWE_Alpha

T3_WavEntr_Alpha

T3_KFD_Alpha

T3_AbsPow_Beta

T3_KFD_Beta

T3_AbsPow_Gamma

T3_KFD_Gamma

C3_AbsPow_Delta

C3_AbsPow_Alpha

C3_KFD_Alpha

C3_RelPow_Beta

C3_RWE_Beta

C3_WavEntr_Beta

C3_AbsPow_Gamma

Cz_AbsPow_Delta

Cz_RelPow_Delta

Cz_WavEntr_Delta

Cz_KFD_Delta

Cz_AbsPow_Theta

Cz_KFD_Theta

Cz_AbsPow_Alpha

Cz_KFD_Alpha

Cz_RelPow_Beta

Cz_KFD_Beta

Cz_AbsPow_Gamma

Cz_KFD_Gamma

C4_AbsPow_Delta

C4_KFD_Delta

C4_AbsPow_Theta

C4_KFD_Theta

C4_AbsPow_Alpha

C4_WavEntr_Alpha

C4_KFD_Alpha

C4_RelPow_Beta

C4_RWE_Beta

C4_WavEntr_Beta

C4_AbsPow_Gamma

T4_AbsPow_Delta

T4_KFD_Delta

T4_AbsPow_Theta

T4_KFD_Theta

T4_AbsPow_Alpha

T4_KFD_Alpha

T4_AbsPow_Beta

T4_KFD_Beta

T4_AbsPow_Gamma

T4_KFD_Gamma

T5_AbsPow_Delta

T5_SpecCent_Theta

T5_RelPow_Alpha

T5_RWE_Alpha

T5_WavEntr_Alpha

T5_SpecCent_Alpha

T5_RelPow_Beta

T5_SpecCent_Beta

T5_AbsPow_Gamma

P3_AbsPow_Alpha

P3_KFD_Alpha

P3_AbsPow_Beta

P3_KFD_Beta

Pz_KFD_Delta

Pz_AbsPow_Theta

Pz_WavEntr_Alpha

Pz_KFD_Alpha

Pz_KFD_Beta

P4_AbsPow_Delta

P4_RelPow_Delta

P4_RWE_Delta

P4_WavEntr_Delta

T6_RWE_Delta

T6_WavEntr_Delta

T6_SpecCent_Theta

T6_SpecCent_Alpha

T6_SpecCent_Beta

O1_AbsPow_Delta

O1_RWE_Delta

O1_WavEntr_Delta

O1_KFD_Delta

O1_AbsPow_Theta

O1_SpecCent_Theta

O1_RelPow_Alpha

O1_RWE_Alpha

O1_SpecCent_Alpha

O1_AbsPow_Beta

O1_KFD_Beta

O1_AbsPow_Gamma

O1_KFD_Gamma

O2_SpecCent_Theta

O2_KFD_Gamma

Abstract

Explainability of Machine Learning Models in Prediction of Affective Disorders

Fani Sentinella-Jerbić

This thesis explores the explainability of using EEG brain recordings to distinguish between healthy individuals and those diagnosed with depression. To identify patterns and predictive biomarkers, a variety of analytical approaches were used, including exploratory data analysis, hypothesis testing, explainable AI-based feature rankings, and feature subset evaluation. Challenges arose from the dataset's size, the high feature-to-instance ratio, and the dependability of methods on different assumptions and architectures, limiting their effectiveness and agreement. Notably, the analysis suggests differences in central tendencies and variances between the two groups, a non-linear relationship between EEG features and depression, and the potential biomarkers being the frontal brain area, left hemisphere, and beta waves when used in isolation, or prefrontal, frontal, temporal, and parietal regions as well as alpha, delta, and gamma waves when used in combination. While definitive conclusions are constrained by dataset and method limitations, the insights gained lay the groundwork for future EEG-based studies aiming to refine depression diagnostics and treatment strategies.

Keywords: explainable artificial intelligence (XAI), electroencephalography (EEG), machine learning, affective disorders, depression

Sažetak

Objašnjivost modela strojnog učenja u predikciji afektivnih poremećaja

Fani Sentinella-Jerbić

Ovaj rad istražio je objašnjivost uporabe EEG snimanja mozga u razlikovanju zdravih osoba i osoba dijagnosticiranih depresijom. Kako bi se otkrili obrasci i prediktivni biomarkeri, korištene su različite analitičke metode, uključujući istraživačku analizu podataka, testiranje hipoteza, rangiranje značajki pomoću metoda objašnjive umjetne inteligencije te evaluaciju podskupova značajki. Mali skup podataka, visok omjer značajki i instanci te ovisnost korištenih metoda o različitim pretpostavkama i arhitekturama, ograničili su učinkovitost i međusobno slaganje među primjenjenim metodama. Ipak, provedene analize sugeriraju razlike u središnjim tendencijama i varijancama između zdravih i depresivnih individua, nelinearni odnos između EEG značajki i depresije, te frontalno područje mozga, lijevu hemisferu i beta valove kao potencijalne biomarkere depresije u slučaju izolirane analize. U slučaju kombinirane analize, predfrontalno, frontalno, temporalno i parijetalno područje, kao i kombinacija alfa, delta i gamma valova pokazuju prediktivnu moć. Iako su konačni zaključci ograničeni prirodom skupa podataka i metodologije, dobiveni uvidi postavljaju temelje za buduća istraživanja temeljena na EEG-u s ciljem unapređenja dijagnostike i strategije liječenja depresije.

Ključne riječi: objašnjiva umjetna inteligencija, elektroencefalografija (EEG), strojno učenje, afektivni poremećaji, depresija