Emotion-Based Melody Generation using Song Vocals and Psychophysiological Signals

L.V.N. Wijethunge

Emotion-Based Melody Generation using Vocals and Psychophysiological Signals

L.V.N Wijethunge
Index No: 20002191

Supervisor: Dr. M.I.E. Wickramasinghe

May 2025

Submitted in partial fulfillment of the requirements of the B.Sc. (Honours) in Computer Science Final Year Project

Declaration

I certify that this dissertation does not incorporate, without acknowledgement, any material previously submitted for a degree or diploma in any university and to the best of my knowledge and belief, it does not contain any material previously published or written by another person or myself except where due reference is made in the text. I also hereby give consent for my dissertation, if accepted, be made available for photocopying and for interlibrary loans, and for the title and abstract to be made available to outside organizations.

Candidate Name : L.V.N. Wijethunge

.....
Signature of Candidate Date: 30th June 2025

This is to certify that this dissertation is based on the work of Mr. L.V.N. Wijethunge under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Principle Supervisor Name: Dr. M.I.E. Wickramasinghe

.....
Signature of Supervisor Date: 30th June 2025

Co-Supervisor Name : Mr. N.H.P.I. Maduranga

.....
Signature of Co-Supervisor Date: 30th June 2025

Acknowledgements

I would like to express my sincere gratitude to my supervisors at COTS Lab, Dr. Manjusri Wickaramasinghe, Mr. Isuru Nanayakkara, and Mr. Roshan Nadeesha, for their invaluable expertise, guidance, and unwavering support throughout the course of this research. Their insights and encouragement were instrumental in shaping my understanding of the subject and enhancing my research skills. I am also deeply grateful to Mr. Pasindu Marasinghe for his initial support and encouragement, which laid the foundation for this work.

Additionally, I would like to extend my heartfelt thanks to my colleagues and friends who assisted in data collection and provided insightful feedback and constructive criticism. Their contributions were essential in refining and improving the quality of this thesis.

Finally, I would like to acknowledge all others who have contributed in any way to the success of this research. Without their collective support, this work would not have been possible.

Abstract

This research presents a novel framework that bridges EEG-based emotion recognition and AI-driven music generation, focusing on emotions evoked by song vocals. While earlier studies have explored music emotion recognition and EEG analysis separately, none of the literature has connected vocal-induced emotional states to generative music systems. This study addresses that gap by proposing a system that interprets arousal and valence values derived from EEG signals while listening to vocal tracks and translates them into emotionally aligned melodies.

The experiment involved 30 participants who listened to 104 carefully curated vocal songs. EEG signals were recorded and processed using advanced feature extraction techniques, including Discrete Wavelet Transform (DWT) and Continuous Wavelet Transform (CWT). These features were input into machine learning models such as Support Vector Regression (SVR), Long Short-Term Memory (LSTM) networks, and a hybrid CNN + LSTM architecture. Results showed that the LSTM model achieved strong predictive accuracy with a Mean Absolute Error of 0.042 for arousal and 0.057 for valence, though it lacked spatial feature representation. The CNN + LSTM model demonstrated superior performance by capturing both spatial and temporal EEG features.

A key novelty of this work lies in its end-to-end pipeline that converts predicted emotional states into natural language prompts, which are then used to condition MUSICGEN, a transformer-based music generation model. This enabled the creation of emotionally congruent music, with user-controllable parameters such as melody duration, instrumentation, genre, and tempo.

Emotion alignment analysis revealed high consistency in arousal based generation, with over 80% alignment for most tracks. Valence alignment, while promising for several tracks (above 90%), exhibited greater variability, highlighting challenges in capturing subjective emotional tones.

The dataset curated during this study has been made publicly available to support future research in affective computing, emotion-aware music generation, and human-computer interaction.

Table of Contents

List of Figures

List of Tables

Acronyms

AMG Automatic Music Generation. 4, 5, 35

CNN+LSTM Convolutional Nueral Network + Long Short-Term Memory. 42, 43, 46, 47, 49

CWT Continuous Wavelet Transform. 41, 49

DES Differential Emotions Scale. 10

DWT Discrete Wavelet Transform. 41, 49

ER Emotion Recognition. 2–8, 31

GEW Geneva Emotion Wheel. 10

LSTM Long Short-Term Memory. 42, 43, 45–47, 49

MG Music Generation. 4

SVR Support Vector Regression. 42, 44, 45, 47

Chapter 1

Introduction

1.1 Music and Emotions

Music has an ability that arouses emotions, tells stories, and unites people (Clarke, DeNora & Vuoskoski 2015). Music serves as an avenue for expression and mood regulation for all mankind. When words cannot adequately express an emotion, music may. A melody is a series of notes that flow smoothly together and are perceived by the listener as one whole. The combination of Pitch and rhythm makes up a melody in the most literal sense. However, the phrase can also refer to other elements like tonal color(Wikipedia n.d.). The connection between lyrics, vocals and melody is what gives music its strength to form a smooth sound. Song lyrics are the words used in a song to convey a message or tell a story. They can be seen as a form of literary work, similar to poetry, and are often used to express emotions or describe a particular situation or mood. Song lyrics can have multiple layers of meaning, including denotation (literal meaning), connotation (symbolic or metaphorical meaning), and social criticism(Baur, Steinmayr & Butz 2010). However, (Krishnan 2023) mentioned in one of his articles that producing songs that accurately convey the emotional meaning of lyrics requires skill and sometimes requires many years, even though some understanding of it may come naturally. A song vocal refers to the human voice in a piece of music, including singing, humming, or vocalizations, which can serve as a central or supporting element. Vocals convey the lyrics and emotional nuances through tone, pitch, dynamics, and expression. They are one of the most powerful components of music for evoking emotions, as they connect directly to human communication and expression.

Emotions are mental and physical states resulting from neurophysiological changes that are correlated by various means, including thoughts, sentiments, activities, and a shared sense of humor or discomfort(Damasio 1998). Emotions, mood, character, personality, attitude, or creativity are linked. The James-Lange theory states that emotions derive from how we perceive bodily changes occur due to external or internal signals(James 1884). This concept states various emotions are associated with particular responses of the body, referring to physical changes that occur due to emotions, such as changes in heart rate, respiration, and skin conductance. These changes are mediated by the autonomic nervous system and can be measured objectively using various physiological signals. Also emotions are also regulated by various area of the brain. In the figure **??** section a depicts the most important brain areas involved in emotional processing, such as emotion generation

and control. Purple boxes indicate areas more associated with emotion generation (Periaqueductal Grey, PAG; amygdala; Nucleus Accumbens, NAcc; striatum ventral). Pink boxes reflect brain areas that have a greater impact on emotion regulation and modulation (ventrolateral, dorsolateral, and dorsomedial prefrontal cortex, vlPFC; dlPFC; dmPFC; inferior parietal region, IFP; supplementary motor area, SMA). Red boxes indicate brain locations that may be involved in both processes (orbitofrontal cortex/vmPFC; insula; anterior cingulate cortex, ACC). Section B shows a more simplified diagram of common brain areas involved in both eating behavior and emotional processing that perform similar roles. The red boxes represent affect modulation, while the blue boxes represent emotion initiation (Godet, Fortier, Bannier, Coquery & Val-Laillet 2022).

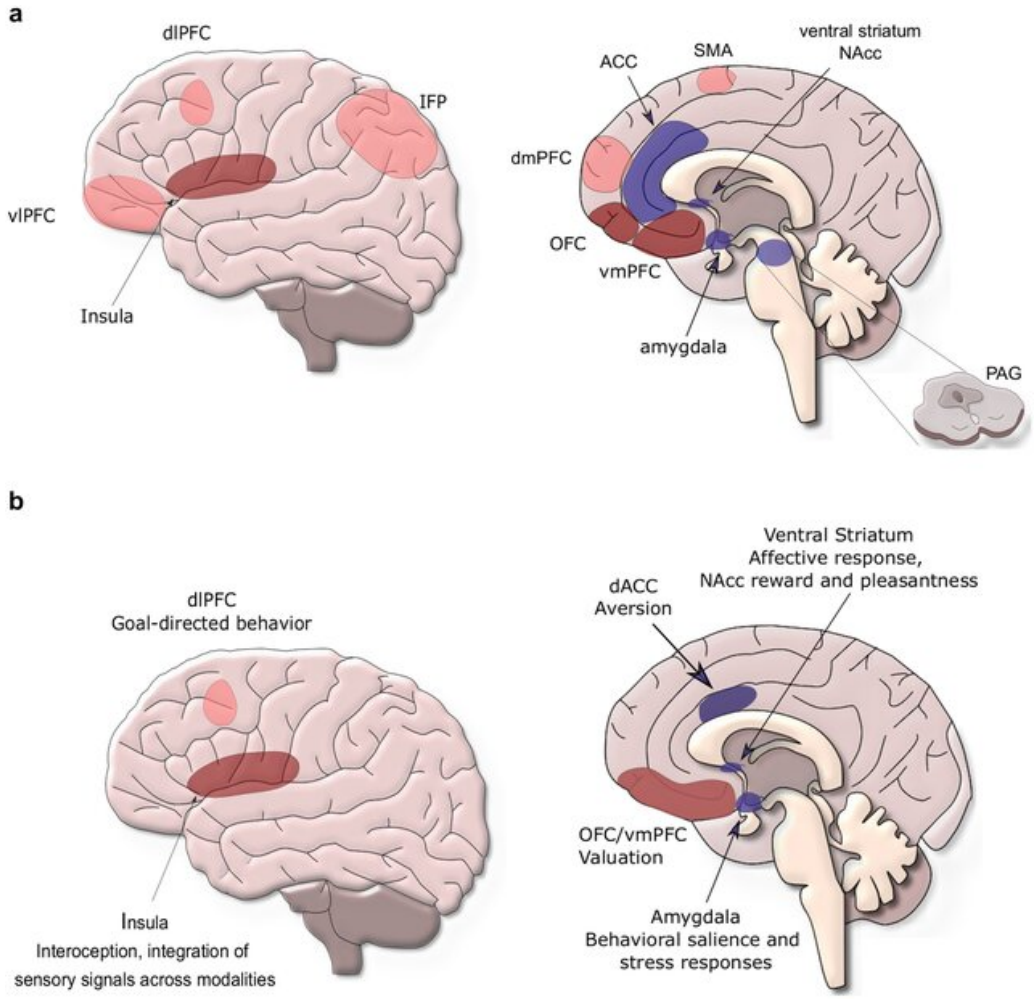


Figure 1.1: Major brain areas involved in emotional processing, including emotion generation and emotion regulation.

1.2 Emotion Recognition

In recent years, Emotion Recognition (ER) has become an important research topic due to its potential applications in various fields such as healthcare, psychology, and marketing. Current methodologies of ER can be broadly classified into two categories(Shown in figures: 1.2 and 1.3):

1. Methods based on Behavioral Responses.

2. Methods based on Physiological Responses.

Figure 1.2: Behavioral Responses

Figure 1.3: Physiological Responses

Behavioral-Based techniques analyze speech, gestures, body language, and facial expressions to identify emotions. Among the techniques in this area that is most frequently employed is facial expression analysis. The focus of this field's current research has been on deep learning-based models that have a high accuracy rate in facial ER(Kollias, Nicolaou, Kotsia, Zhao & Zafeiriou 2017).

However, researchers employ many techniques in physiologically based ways to measure physical signals.

- EEG (Electroencephalography) for brain activity
- ECG (Electrocardiography) for heart activity
- EOG (Electrooculography) for eye movement
- EMG (Electromyography) for muscle activity

- MRI (Magnetic Resonance Imaging) for blood flow in the brain

are some examples of these signals. Since physiological-based procedures are unaffected by environmental circumstances, cultural background, or subjective willfulness, they are seen as more objective. Furthermore, because they can yield more precise measurements of emotions, physiologically based procedures are significant. For example, physiological approaches can offer a better understanding of patients' emotional responses than behavioral approaches when it comes to specific neurological illnesses, including autism spectrum disorder (Fan, Yan, Xiaomin, Yan, Li, Xie & Yin 2020).

1.3 EEG-Based Emotion Recognition

EEG-based ER has drawn the most attention among these techniques because of its high temporal resolution and non-invasive nature. EEG can reveal information about a person's emotional state by measuring the electrical activity of the brain. Many domains, such as marketing, healthcare, and human-computer interaction (HCI), have made use of EEG-based techniques. Customers' emotional reactions, for instance, have been measured using this method to assess the efficacy of marketing efforts (Vecchiato, Maglione, Cherubino, Wasikowska, Wawrzyniak, Latuszynska, Latuszynska, Nermend, Graziani, Leucci, Trettel & Babiloni 2014). Within the medical field, EEG-based ER (ER) has been utilized to track the emotional states of patients with mental illnesses and evaluate the effectiveness of therapy (Nuri, Niazi & Guger 2019).

Research on emotion detection relies heavily on stimuli since they offer a means of evoking quantifiable and analyzed emotional reactions. Studies that seek to understand the brain mechanisms behind emotional processing should make special use of stimuli because they enable the controlled manipulation of emotional states (Kreibig 2010).

1.4 Music Generation

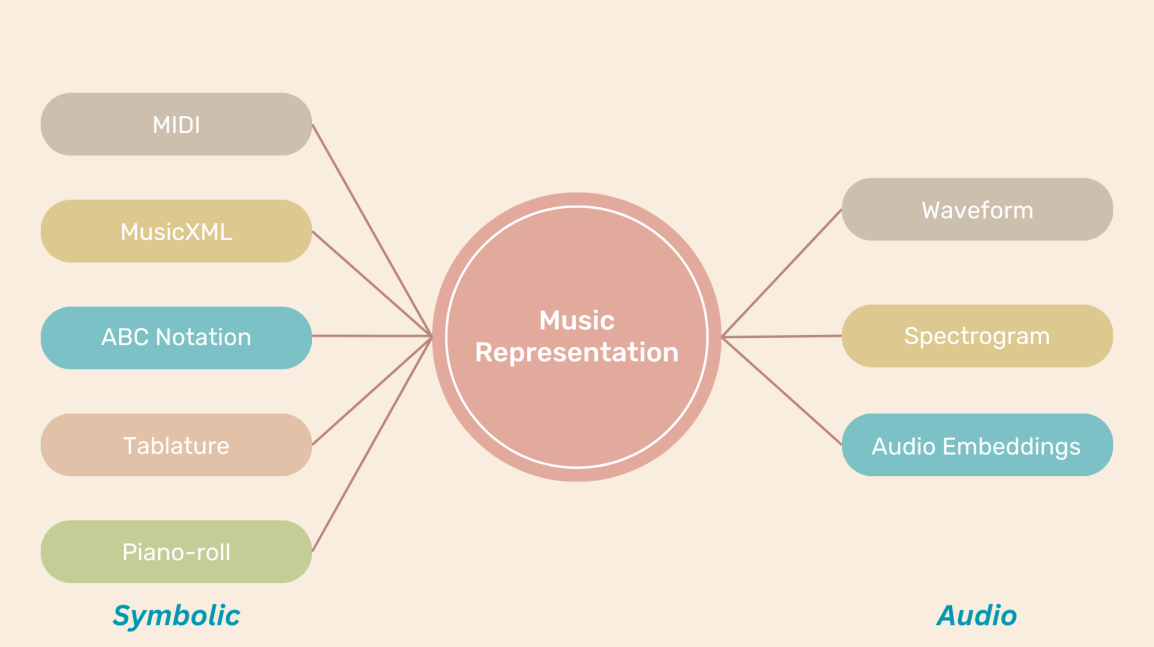
Music Generation (MG) can be know as generating sounds or music from a model or algorithm. The goal is to produce a sequence of notes or sound events that are similar to existing music in some way, such as having the same style, genre, or mood. MG possibilities have been opened up by the recent advancements in artificial intelligence and machine learning. One of the eye-catching definition was for Generative Music was "The art and science of developing computer programs that create music with a varying degree of autonomy" by (ValerioVelardo 2023). Creating music with a varying degree of autonomy explains how successful the automation of the various systems that generate music has been. Symbolic music generation and audio music generation are two categories of Automatic Music Generation (AMG) techniques (Guo, Liu, Zhou, Xu & Zhang 2023). The history of generative music dates back to the 1700s but gained widespread popularity in recent years. (ValerioVelardo 2023) outlines the evolution of generative music through five distinct eras. The Pre-computing Era (1700-1956) featured manual algorithms like Mozarts Dice Game. The Academic Era (1956-2009) were computer-based compositions, such as the Illiac Suite. From 2010-2016, the first startup wave emerged, introducing machine learning and AI music generation tools. The Big Tech Experiments Era (2016-2022) focused on deep learning, with notable advancements like Amazons Deep Composer and OpenAIs JukeBox. By the late 2020s, generative AI exploded, with models like Googles

MusicLM and Meta AIs MusicGen leading the Music AI Hype. These advancements have revolutionized the field, making AMG increasingly scalable and accessible. All of the music that has been generated by AI comes under two representations,

- Symbolic Representation
- Audio Representations

Example representation types can be seen in the figure 1.3.

Figure 1.4: Types of Music Representation



1.5 Background to the Research

ER and music generation have been an interesting research area in the past decade, and people tend to listen to music that gives some kind of personalization to their emotions. Something that resonates with their personal feelings. There has been some attention to generate melodies related to your emotions that had the capacity to evoke, amplify, or modulate emotions, making it a powerful tool for emotional expression and regulation. One of the primary focuses of this research area is to explore on how can music match with specific emotional responses, often leveraging techniques in machine learning, natural language processing, sentiment analysis techniques and affective computing. People tend to select music that resonates with their emotional state, seeking songs that reflect or alter their feelings, which has spurred growing interest in the personalization of music. A previous study (Wijethunge, Akarawita, Hegodaarachchi, Abeytunge, Gamage & Wickramasinghe 2024) focused on generating melodies based on music likability to introduce an element of personalization. However, recent observations indicate that songs featuring vocals often evoke a stronger emotional resonance with listeners. People feel a deeper emotional connection to music with vocals compared to purely instrumental songs. It's like there's something special about hearing actual human voices that touches our hearts in a way that just instruments can't quite manage.

When someone sings, it's not just about the words - it's how they say those words. The way a singer's voice rises and falls, the warmth or roughness in their tone, the emotions that slip through between the lyrics. All these create a much more powerful emotional experience. Think about how a sad song can make you feel the pain in the singer's voice, or how a love song can make you feel the passion through their vocal performance. Lyrics play a big part too. Words give context and meaning that pure instrumental music can't easily communicate. When someone sings about heartbreak, joy, or struggle, it feels more personal and relatable. It's like the singer is telling a story directly to you, sharing an emotional journey that you can connect with. This study wants to explore this deep emotional connection by looking to create a way to generate music that captures these vocal emotional elements. Instead of just creating melodies, wanted to understand how to make music that feels personal, that sounds like it's speaking directly to a listener's heart and experiences.

The goal is to go beyond just creating attractive music. Focues on how to create melodies that truly resonate with people's emotions, using the power of vocal expression to create a more meaningful musical experience.

1.6 Problem Definition

ER from EEG signals is a challenging task due to the complexity of brain activity and the influence of individual and external factors. Conventional EEG-based ER frameworks involve several steps, including stimuli selection, EEG data collection, pre-processing, feature extraction, and classification. However, existing methods face challenges such as noise in EEG signals, variability in emotional responses, and the difficulty of accurately mapping EEG features to emotional states.

Current approaches often rely on predefined datasets with static annotations, which may not fully capture the dynamic and subjective nature of emotions, particularly in response to auditory stimuli like song vocals. In addition, selecting the most relevant EEG features and designing an effective model for emotion classification remain open challenges.

This research aims to improve EEG-based ER by developing a novel framework that accurately maps EEG signals to emotional values (arousal and valence) when a person listens to song vocals. By optimizing EEG preprocessing techniques, refining feature extraction methods, and employing machine learning models tailored for emotion classification, this study seeks to enhance the accuracy and reliability of EEG-based emotion detection. The resulting emotional values will serve as the foundation for music that aligns with the emotional state of the listener.

1.7 Research Aim, Questions, and Objectives

1.7.1 Research Aim

To build an EEG-based ER framework with improved accuracy and efficiency in detecting emotional states from song vocals. The extracted emotional values will be used to generate melodies, thus creating a personalized psychological connection between the listener and the music.

1.7.2 Research Questions

RQ1 *What is the most suitable methodology for accurately identifying and quantifying emotional responses expressed by song vocals?*

Identifying the most suitable methodology for ER from song vocals involves evaluating various physiological and computational approaches. This includes analyzing techniques like EEG-based modeling, acoustic analysis, or multimodal fusion methods. This research question aims to determine which approach or combination of approaches offers the most accurate, efficient, and scalable solution for quantifying emotional responses, considering both the emotional richness of vocals and the complexity of human affective states.

RQ2 *What is the proper systematic approach that will give a likely balanced and representative dataset for the selected methodology?*

The second research question aims to address the critical need for a well-structured and representative dataset to support the chosen ER methodology. Emotional data, especially from sources like EEG or audio, often suffer from imbalance, noise, or subjectivity. Establishing a systematic approach involves selecting appropriate stimuli (e.g., song vocals), consistent labeling strategies (e.g., using arousal-valence models), and segmenting data to ensure temporal uniformity. This question seeks to explore preprocessing techniques, annotation strategies, and data balancing methods that contribute to a dataset capable of training reliable and generalizable models.

RQ3 *What is the impact of different feature extraction methods and model architectures on the accuracy and robustness of selected ER method?*

Feature extraction and model architecture are central to the performance of any ER system. This research question investigates how various signal processing techniques such as Fourier Transform, Discrete Wavelet Transform (DWT), statistical descriptors, or frequency domain analyses affect the quality of emotional features extracted from the input data. In parallel, it examines how deep learning architectures like LSTM, CNN, or hybrid CNN-LSTM models handle these features in terms of classification accuracy and robustness. The goal is to identify optimal combinations that maximize performance while minimizing overfitting and computational complexity.

RQ4 *How can the emotional values derived from EEG responses be meaningfully mapped into a generative music model?*

The final research question focuses on the translation of quantified emotional states (e.g., arousal and valence scores) into a format that can guide a generative music model. Given that generative models like MusicGen require textual or symbolic inputs, this step involves designing a meaningful mapping process from numerical emotional data to descriptive emotional prompts. This ensures that the emotional intent captured from the EEG responses is preserved in the resulting music. The challenge lies in aligning the emotional semantics of brain-derived data with the expressive capacity of AI-based music generation tools.

1.7.3 Research Objectives

RO1 *To explore and evaluate different methodologies for detecting emotional responses to song vocals.*

This objective focuses on comparing physiological and computational approaches such as EEG analysis, acoustic ER, or statistical approaches to identify the most effective framework for emotion detection.

RO2 *To develop a systematic data collection and pre-processing pipeline to construct a balanced and representative dataset.*

This includes selecting suitable data acquisition methods, annotation methods (e.g., arousal-valence labeling) and effective data pre-processing methods to reduce bias and noise.

RO3 *To extract meaningful emotional features using advanced signal processing techniques.*

This involves implementing and comparing Continuous Wavelet Transform(CWT), , Fast Fourier Transform(FFT), and statistical measures to derive features that reflect emotional patterns in EEG or vocal signals.

RO4 *To design and evaluate deep learning models for classifying emotional states.*

The goal is to apply architectures like LSTM, CNN, and hybrid models, assessing their performance in terms of accuracy, generalization, and computational efficiency.

RO5 *To map predicted emotional values into descriptive text prompts for music generation.*

This objective aims to translate the arousal valence output into emotion-rich textual input compatible with models such as MusicLM, MusicGen allowing personalized melody generation.

1.8 Justification of the Research

Emotions are fundamental to being human. They determine what we perceive, decide, and relate to one another. Utilizing brain-computer interface (BCI) technologies such as EEG to recognize emotions (ER) is very fashionable now, particularly for web-based and non-invasive applications. Though significant progress has been made, there are still technical challenges in actually interpreting and deciphering emotional responses, particularly for something like music vocals, which can elicit strong and varied emotions. ER systems today do not typically accommodate personal needs and do not encompass the entire scope of emotional expression articulated in song vocals.

This research addresses the gap of needing a more effective system that identifies emotions and take some initial steps for the use of EEG analysis with music technology. It demonstrates how vocal songs affect emotions and connects emotional values from EEG to music generation. The objective is to enhance ER using better preparation methods, improved methods of identifying key features, and evaluation of deep learning models tailored specifically for emotions elicited by singing.

Chapter 2

Literature Review

The literature review was under two main sections as shown in figure **??**.

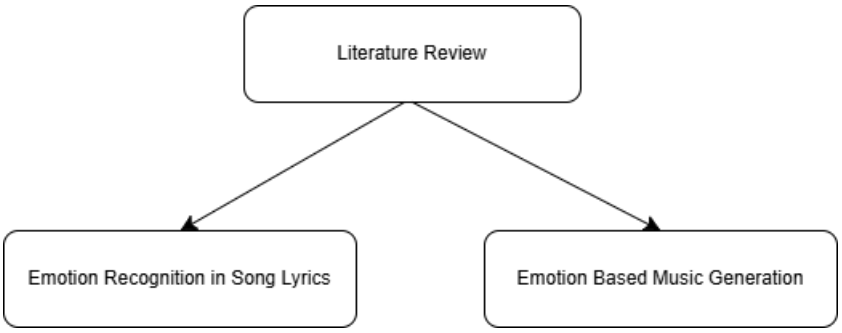


Figure 2.1: Literature Review Taxonomy

2.1 Emotion Models

Emotions are complicated feelings composed of combinations of thoughts, bodily reactions, and behaviors. They arise due to what we feel internally or due to what is taking place externally, and they influence most human behaviors such as decision-making, remembering, and communication with other people. Emotions play a crucial role in psychology and neuroscience, and also in fields such as emotional computing and human-computer interaction.

Emotions have been explained in various ways from a theoretical perspective. The James-Lange theory of emotion implies that emotions emerge as a consequence of changes in the body due to external stimulation. The brain captures the changes and interprets them as particular emotions. Contrarily, the Cannon-Bard theory of emotion implies that feelings of emotion and bodily reactions occur simultaneously and in parallel (Levenson 2014). Current research integrates these theories by acknowledging the role of brain systems in the production of emotions, emphasizing the coordination of brain regions such as the amygdala, prefrontal cortex, anterior cingulate cortex, and insula (Pessoa 2013); (Craig 2009). These regions are involved in the detection of emotional cues, regulation of emotional responses, and monitoring of bodily states associated with excitement (Phelps 2006); (Shackman, Salomons, Slagter, Fox, Winter & Davidson 2011).

Researchers in this study area have proposed various models. Two main models are Categorical models and Dimensional models.

2.1.1 Categorical Models

Categorical models explain emotions as distinct and identifiable categories such as anger, fear, happiness, and sadness. These models are often applied with instruments such as the Geneva Emotion Wheel (GEW) and the Differential Emotions Scale (DES). The Geneva Emotion Wheel (GEW) classifies 20 categories of emotions into four categories depending on whether they are positive or negative and how energetic they are (Shuman, Schlegel & Scherer 2015). The Differential Emotions Scale (DES) contains 15 emotions based on basic feelings (Juslin & Västfjäll 2008)(Figure 2.2). These models are simple to apply and interpret and, therefore, are often applied in psychological research and computer programs that recognize emotions.

Figure 2.2: Differential Emotions Scale (DES)

Table 10.1. *The differential emotions scale*

Factor	Item	Item–factor correction
I. Interest (.76)	attentive	.88
	concentrating	.79
	alert	.87
II. Enjoyment (.87)	delighted	.81
	happy	.87
	joyful	.86
III. Surprise (.75)	surprise	.83
	amazed	.85
	astonished	.87
IV. Sadness (.85)	downhearted	.86
	sad	.79
	discouraged	.82
V. Anger (.68)	enraged	.74
	angry	.84
	mad	.86
VI. Disgust (.73)	feeling of distaste	.86
	disgusted	.85
	feeling revulsion	.78
VII. Contempt (.78)	contemptuous	.89
	scornful	.90
	disdainful	.84
VIII. Fear (.68)	scared	.88
	fearful	.90
	afraid	.89
IX. Shame/shyness (.83)	sheepish	.73
	bashful	.87
	shy	.88
X. Guilt (.77)	repentant	.78
	guilty	.83
	blameworthy	.80

Note: Item–factor correlations for state instructions, $N = 259$; test–retest reliabilities for trait instructions given in parentheses, $N = 63$.
Source: Izard, 1977.

But categorical models have their limits. First, they oversimplify the emotion, even though emotions are variable and complex. Individuals feel emotions as intensities or mixtures, not as fixed states, and therefore, it is difficult to classify them into fixed states. Second, individuals and cultures read and apply emotions differently, and it becomes more difficult to apply these categorical models across the board. Third, the requirement to respond in fixed categories when describing feelings can limit the things that individuals can report, and therefore, their descriptions of emotion become less valid and informative (Moncrieff & Lienard 2018).

2.1.2 Dimensional Models

To counter these shortcomings, dimensional models have become increasingly popular. Such models depict emotions as scales rather than categories. The most common dimensional model is the Valence-Arousal model proposed by (Russell & Barrett 1999), in which emotional states are placed in two-dimensional space: valence (positive to negative) and arousal (high to low intensity). Certain models even have a third dimension, dominance, to encompass the perception of control of an emotional state.

In Russell's 2D valence-arousal dimensional emotion space(Figure 2.3), when considering the emotion of happiness, it must be located in the top first quadrat as it is considered to be a positive emotion (valence) that has a high intensity (arousal). The same for sadness, it should be in the bottom left quadrant as it is considered to be a negative emotion with lower levels of arousal.

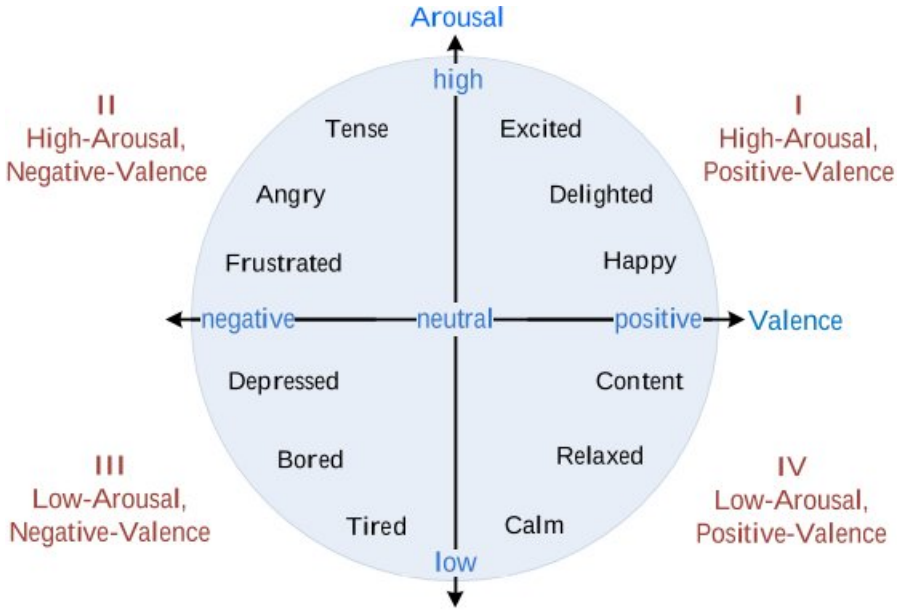
Dimensional models are more adaptable and better capture the subjective and multifaceted nature of emotions. Dimensional models are also consistent across different cultures and people by allowing labeling of emotions using a universal system rather than culturally labeled terms (Moncrieff & Lienard 2018). Dimensional models in emotion recognition research using EEG are particularly beneficial because they align well with continuous brain signal data and allow for real-time monitoring of emotions.

One of greates virtues of Russell's model is its simplicity and explanatory power. By positioning emotions on two significant dimensions, the model provides you with a simple means of comparing and quantifying emotional experience. (Juslin & Västfjäll 2008) describe arousal as how active or intense an emotion is and valence as how positive or negative it is. These dimensions relate directly to simple behaviors such as approach or avoidance, which are central to the way we feel and decide. (Russell & Barrett 1999) went as far as suggesting that arousal and valence are the "core processes" of emotion, the basic feelings behind more complex emotional states.

Although this model is helpful, it has been criticized somewhat. A few scientists propose that the consideration of emotions in only two dimensions may render it too simplistic to understand and may overlook significant differences between various emotions [Lazarus, 1991]. For instance, anger and fear can have the same degree of excitement and negative emotions, but very distinct in what we think about them, how we behave, and what provokes them. Critics propose that such differences may get confused in a straightforward dimensional model, which could render emotion categorization less accurate ((Russell & Barrett 1999); (Mauss & Robinson 2009). The 2D valence-arousal model continues to be useful when analyzing emotions across different domains including music, film, commercials, and cross-cultural research (Gomez, Danuser & Grimm 2019). Happiness, with high valence and high arousal, for example, will usually be positioned in the upper-right quadrant of the DES, and sadness, with low valence and low arousal, in the lower-left quadrant. These positions are used to clearly associate emotional reactions with different stimuli and groups.

In EEG emotion recognition, both categorical and dimensional theories have been used by researchers. Categorical theories are easier to annotate and easier to classify but are prone to confusion and overgenerality, particularly with emotions that are continuum-based or difficult to categorize neatly. For instance, (Juslin & Västfjäll 2008) employed categorical emotion labels in an EEG experiment and achieved a 70% accuracy.

Figure 2.3: Russells's Arousal Valence 2D Model



Dimensional models provide a more nuanced account of emotions and are better suited to the continuous nature of physiological signals such as EEG. These are illustrated by (Zheng 2015) with a capacity to achieve a 77.1% recognition rate by using the valence-arousal dimensional model. But one of the problems is that it is more cognitively taxing on annotators since comprehension and use of dimensional ratings require greater levels of abstract thinking and knowledge of the scale. This proves difficult for inexperienced participants and can lead to inconsistency in marking emotions.

In short, there are two primary approaches to modeling emotions, categorical and dimensional. Categorical models are intuitive and easy to interpret, but they may miss some information and not be optimal in every context. Dimensional models, such as the valence-arousal model, represent emotions in a more fine-grained and continuous manner, but they can be complicated and difficult to apply, particularly in large EEG studies. Achieving the right balance between being easy to interpret and descriptive is one of the greatest challenges in emotion recognition. Future work must consider the trade-offs seriously when choosing an emotional representation framework, particularly for EEG studies with real-world objects such as music.

2.2 Emotion Recognition

Emotion Recognition study was initiated by Charles Darwin. Darwin suggested that emotions are involved in evolution, are entrenched in human biology, and have evolved in order to survive and communicate (Darwin 1872). His idea sensitized us to the fact that emotional signals are universal signs conveyed by biology.

The area of automated emotion recognition started to emerge with new computer technologies. One such breakthrough was offered by Rosalind Picard, whose valuable contribution in Affective Computing (Picard 1997) put the idea of machines sensing and reacting to human emotions in the spotlight. Her initial work, especially in creating systems that can detect subtle facial expressions through computer vision, helped in the

creation of emotion recognition technologies in various fields like healthcare, marketing, and human-computer interaction (HCI).

Scientists have questioned numerous alternative ways of emotion recognition (ER) from various signals and behaviors throughout history. One of these areas is facial expression analysis, which received a great deal of attention. Paul Ekman and Wallace Friesen were the leaders in this area. Their Facial Action Coding System (FACS) (Ekman & Friesen 1978) introduced a scientific approach to the study of facial muscle movement. This system enabled objective measurement of emotional expressions and became common in psychological and computer-based ER research. In addition to facial expressions, other methods have been investigated, such as speech, body cues (e.g., heart rate, skin conductance, and brain waves), and behavioral patterns. For instance, emotion recognition from speech has been extensively researched, e.g., (Schuller 2013) in their research on acoustic parameters and language cues. Similarly, (Busso, Deng, Yildirim, Bulut, Lee, Kazemzadeh, Lee, Neumann & Narayanan 2004) highlighted the importance of combining various types of inputs to develop more robust emotion recognition systems.

The latest advancements in deep learning (DL) and machine learning (ML) have revolutionized how we comprehend emotion recognition. Techniques founded on deep neural networks (DNNs) have proved extremely accurate in identifying complex patterns in detailed emotional data. The yearly Emotion Recognition in the Wild (EmotiW) competition, initiated in 2013, has become a benchmark for evaluating ER systems in practical applications. The competition generally involves the state-of-the-art systems based on Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which are efficient in processing spatial and time-based information respectively (Zeng 2019);(Zhang 2020).

The application of deep learning in affective research has enabled researchers to cope with varied and evolving affective data. The technology is taking the field to real-time applications that take context into account in education, entertainment, and mental health measurement.

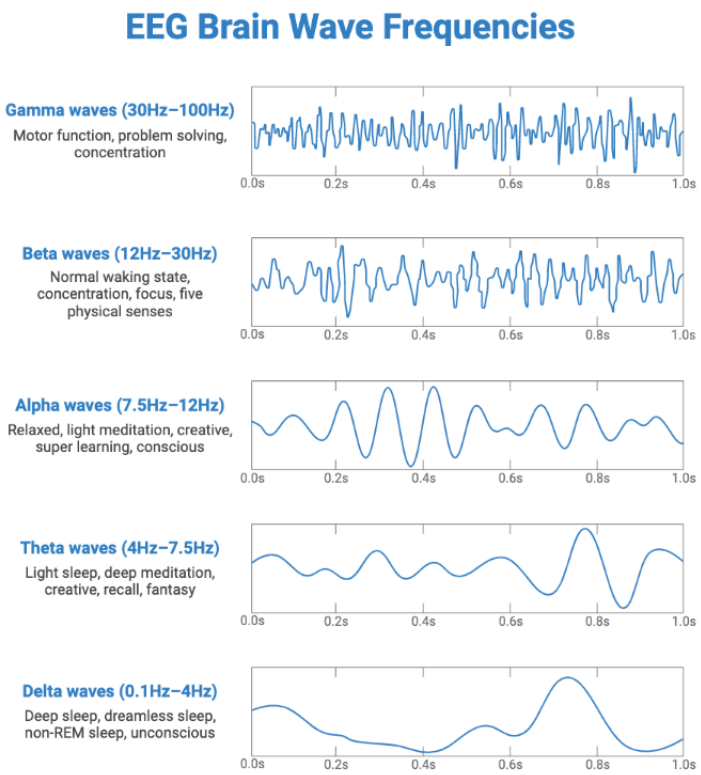
2.3 EEG-based Emotion Recognition

The electroencephalogram (EEG) is a non-invasive neuroimaging method that measures the electrical activity of the brain by applying electrodes to the scalp. EEG technology began with Hans Berger's first experiments in the 1920s. Berger initially measured electrical activity from the human brain and stored these signals in the form of rhythmic wave patterns, the beginning of modern EEG (Berger 1929). Berger's technique was to apply electrodes to the scalp and measure electrical signals using a galvanometer. EEG systems have improved considerably over the years and now possess multi-channel arrays with the ability to measure brain activity with millisecond-level temporal resolution (Davidson 2002)

EEG is based on the premise that brain cells talk to one another by creating voltage changes. These changes appear as changing wave patterns on the scalp. The wave patterns are classified according to their frequency and amplitude into typical bands: delta (0.5 - 4 Hz), theta (4 - 8 Hz), alpha (8 - 13 Hz), beta (13 - 30 Hz), and gamma (>= 30 Hz). Each band is associated with various states of thinking and behavior. For

instance, alpha and beta waves are associated with relaxation and alertness, whereas delta and theta are associated with sleep and deep relaxation. The gamma band is generally associated with processing sensory information and awareness (Yasin, Hussain, Aslan, Raza, Muzammel & Othmani 2021). Figure 2.4 shows the EEG Brainwave frequency and their brain states.

Figure 2.4: EEG Frequency Bands



EEG-based emotion recognition relies on the fact that our emotions cause diverse brain activity patterns. Certain brain regions, such as the amygdala, insula, and prefrontal cortex, are involved in emotion processing and are associated with EEG signal changes.

Initial research indicated that positive affect is associated with increased alpha and beta activity in the left prefrontal cortex. Negative affect, on the other hand, is found to induce delta and theta activity in the right prefrontal cortex (Davidson 1990);(Davidson 2002). Furthermore, emotions such as fear and anxiety are associated with high gamma activity in the amygdala. This indicates the significance of high-frequency waves in emotional arousal (Adolphs 2002); (Herrmann, Matthias & Andreas 2005).

Other emotion-specific patterns have also been found:

- Happiness is associated with greater alpha power in the left front region of the brain (Davidson 1990).
- Sadness is associated with higher theta activity in the right prefrontal cortex (Schmidt & Trainor 2001).

- Anger is related to heightened beta and gamma activity in the prefrontal and anterior cingulate regions (Knyazev 2013); (Gable & Harmon-Jones 2010).

The study of emotion identification with EEG began in the late 1970s and early 1980s. EEG activity was recorded by researchers as subjects viewed emotional stimuli such as pictures, sounds, or videos. Early work revealed that different EEG wave patterns corresponded to different emotions. Alpha waves corresponded to relaxation, and beta waves corresponded to heightened alertness (Gunes & Piccardi 2011). Problems of interpreting the signals and distinguishing fine emotions slowed early advances.

Subsequent work considered extracting significant features, attempting to transform raw EEG data into useful information to distinguish between emotions. Machine learning (ML) significantly enhanced this capability. Early models employed Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) to distinguish emotional states from EEG (Keil, Bradley, Hauk, Rockstroh, Elbert & Lang 2002), but they did not succeed because of high inter-subject variability as well as the complex nature of EEG signals.

In recent years, DL has facilitated the amazing progress in EEG emotion recognition. Network types such as CNNs and RNNs have been shown to possess significant capacity in emotion identification through the automatic learning of meaningful features from raw EEG input (Acharya, Oh, Hagiwara, Tan, Adeli & Subha 2018). These models enable correct classification, capturing subtle patterns that earlier methods used to miss.

Some major findings of current EEG-based emotion research are:

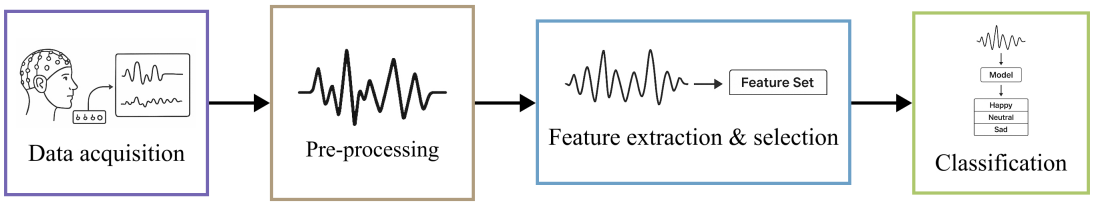
- Frontal Alpha Asymmetry (FAA) is a consistent indicator of approach-avoidance behaviors (Coan & Allen 2004).
- Individual differences in gender, age, and culture influence EEG responses to emotional stimuli. This implies that we require individualized models for emotion detection (Cartocci, Modica, Rossi, Inguscio, Arico, Levy, Mancini, Cherubino & Babiloni 2019).

Cumulatively, these developments highlight the increasing promise of EEG as a neurophysiological instrument for recording and interpreting human emotional states, with potential applications in mental health, adaptive systems, and affective computing.

2.4 Framework of EEG-based Emotion Recognition

Framework for EEG-based emotion recognition can be introduced under five steps(Figure: 2.5. (S. Alarcao and M. J. Fonseca 2019)

Figure 2.5: Steps of EEG-based Emotion Recognition Framework



2.4.1 Data Acquisition

The data acquisition phase is the first and most important aspect of EEG-based emotion recognition. Research requires thoughtful consideration to obtain true and correct results. In this section, you can see what are the steps you should follow in the data acquisition phase, including participant recruitment, consent process, EEG devices, electrode placement, playing stimuli, annotation process, and recording EEG.

- **Participant Recruitment**

  Finding subjects is an important part of research that uses EEG to recognize emotions. Researchers got subjects from different places, like university students and hospitals. patients and community members. The process of selection must be random and fair. to ensure that no bias in the data set exists. The population sample should be of a significant size to depict the population and facilitate statistical analysis. Various studies have used various methods to recruit participants, for example, advertisements, leaflets, and word of mouth. For instance, (Hamedi et al. 2020) enrolled 40 healthy male and female volunteers through ads on social media and bulletin boards, while (Yuchen Zhang et al. 2020) recruited 15 healthy male volunteers through flyers and word-of-mouth.

- **Consent Process**

  Informed consent is a significant and necessary element of the research process because it ensures that the participants know the aim of the research, the hazards, and their rights as participants. In past studies, consent was provided in written form, and the subject was allowed time to read and comprehend the consent form prior to signature. The consent form must provide for the subject's data confidentiality and their right to withdraw at any time during the research. In some research, written informed permission of all subjects, and informing subjects of the purpose of the study, the risks involved, and the rights of the participants to withdraw from the study at any time Baur al. 2019; Sander Koelstra and others 2012.

- **Selecting Stimuli**

  Stimuli selection is an important process in EEG-based emotion recognition studies since it can influence the quality of data gathered. The stimuli should be selected according to their efficacy for the subjects and their valence and arousal ratings. Application Biased stimuli may produce outcomes that are not valid or reliable, and this can affect the understanding of the outcomes. Moreover, most data sets employed in previous research do not have balanced data, which will most probably influence the validity and generalizability of results. One of the most popular databases employed in studies of emotion recognition from EEG. is the International Affective Picture System (IAPS) database. It is a standardized photos that include various feelings and levels of excitement, and has been employed in numerous research to evoke the emotions of individuals. However, some research has challenged the issue of use of IAPS images, since they might not elicit emotion in every subject (Lang, Bradley, and Cuthbert 2005; Palomba et al. 2000). Other databases

employed Studies on emotion recognition with EEG involve the Geneva Affective Picture Database. (GAPED), Chinese Affective Picture System (CAPS), and Affective Norms for English Words (ANEW) database. However, these databases also come with some limitations, such as not encompassing all affect categories, having unbalanced data, or having non-standardized ratings across cultures (Sander Koelstra et al. 2012).

- **EEG Devices Electrode Types**

  Different EEG devices are employed, with different hardware specifications, different software features, and varying from the number of channels such as Emotiv (14 channels or 32 channels) to higher-end configurations. The nature of the configuration also affects data quality:

  - Wet electrodes provide higher quality signals, but are less comfortable.
  - Dry electrodes are less accurate but more convenient.
  - Semi-dry electrodes are an excellent trade-off between comfort and clean signals.

- **Presentation of Stimuli and EEG Recording**

  The standard procedure includes the installation of the EEG system, the presentation of stimuli (images, sounds, or videos), and the recording of EEG data. The subjects must be restrained to minimize movement or eye blink artifacts (Makeig et al., 2004). Standardized presentation software is used to synchronize stimuli and EEG recordings effectively.

- **Annotation and Labeling**

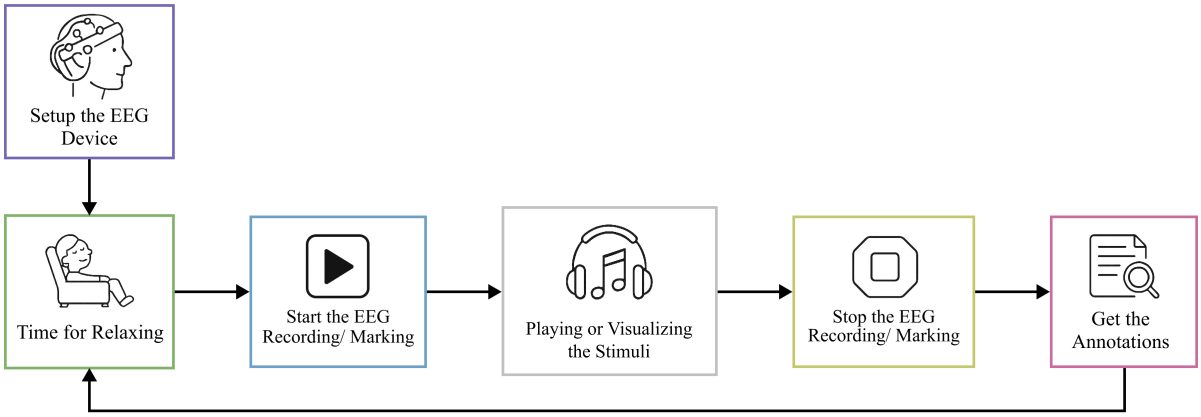
  Following data acquisition, EEG epochs are labeled based on the emotional state of the subject using categorical (e.g., happy, sad) or dimensional (e.g., valence, arousal) models. Annotation is time-consumingZhang et al. (2020) estimated labeling at around 30 hours per participantmaking automation essential.

The overall process of EEG Data Acquisition experiment is shown in figure 2.6. Stimuli selection and annotation are necessary but are under-developed areas in EEG-based emotion recognition. Current approaches are non-standardized and inefficient. Therefore, further research and development are necessary to improve these processes, and it is crucial to explore new and innovative approaches to overcome the current challenges.

2.4.2 Preprocessing

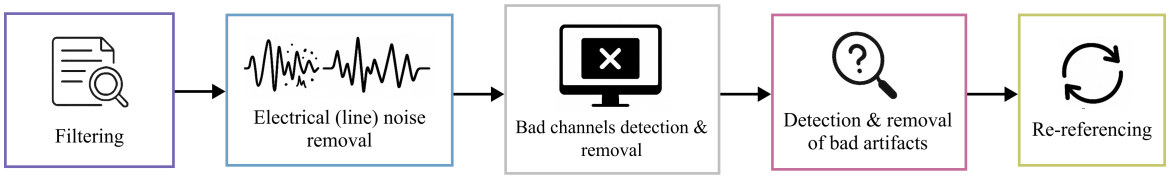
EEG signals are afflicted with a number of types of noise, such as electrical (line) noise, Muscle movement, eye blinking, and other actions can significantly influence how precise things are and the accuracy of emotion recognition outcomes. Hence, a proper preprocessing pipeline is needed to remove these sources of noise and problems and to increase the signal-to-noise ratio of the EEG data. The pre-processing process typically includes

Figure 2.6: Overall process of EEG Data Acquisition experiment



some significant steps, including filtering, electrical (line) noise removal, bad channel rejection, bad artifact detection and removal, and re-referencing. Figure 2.7 illustrates the general pre-processing Pipeline for EEG processing.

Figure 2.7: EEG Pre-processing Steps



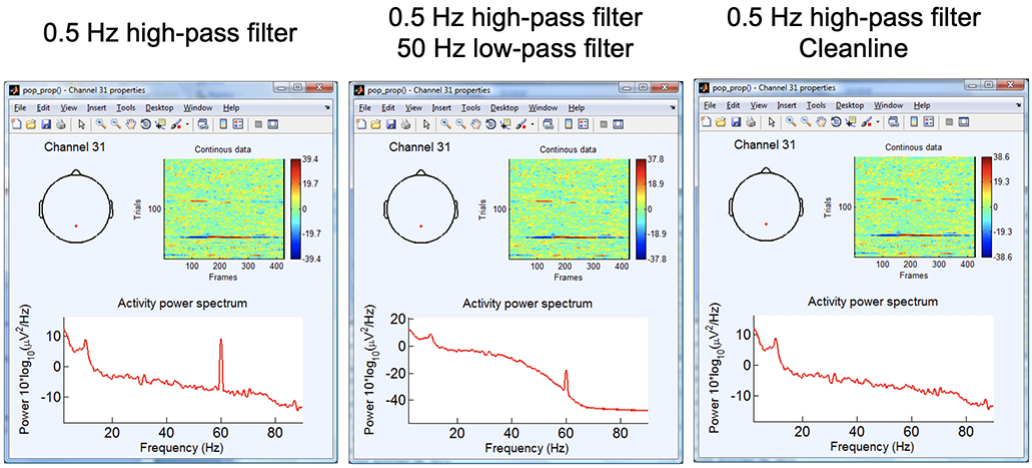
- **Filtering**

  Filtering is an essential pre-processing process in EEG-based emotion recognition research. The primary function of filtering is to eliminate noise from the recorded EEG signals and enhance the critical frequency bands associated with how we experience emotions. EEG signals are frequently infested with a wide range of noise, both biological and environmental noise, and motion artifacts that can impact the validity and reliability of the emotion recognition system (Arvaneh, Gargiulo, and K.-Y. Kim 2018).

  In EEG-based emotion recognition studies, two types of filters are usually utilized: high-pass and low-pass filters. High-pass filters eliminate the low-frequency components of the EEG signals, and low-pass filters eliminate the high-frequency components. A bandpass filter(Figure 2.8) is typically employed to obtain the desired frequency range (Jalilifar and Roshani 2018). The cut-off frequencies of the filters rely on the frequency range of Interest and the character of the EEG signal.

  Also, researchers have proposed some equations and methods for filtering EEG signals. The most commonly used filters include Butterworth, Chebyshev, and Elliptic filters. Butterworth filters are the most widely used filters for processing

Figure 2.8: Bandpass Filtering Process



EEG signals as they are flat. Steep transition from stopband to passband with frequency response. Chebyshev filters provide a sharper transition from stopband to passband but at the cost of passband ripple. Elliptic filters have a steeper transition than Chebyshev filters but with ripples both in the passband and stopband Kumar, Aggarwal, and Singh 2020.

The Butterworth filter is used because it is simple and has a smooth frequency It is a low-pass filter that lowers the intensity of all frequencies greater than a cut-off. frequency. The Butterworth filter transfer function is given by,

$$H(s) = \frac{1}{1 + (\frac{s}{w_c})^{2n}}$$

where s is the complex frequency variable, c is the cut-off frequency, and n is the order of the filter. The Butterworth filter's frequency response is given by,

$$H(f) = \frac{1}{1 + (\frac{f}{w_c})^{2n}}$$

where f is the frequency variable. (Butterworth 1930). Butterworth filters possess a flat the passband frequency response and the slow roll-off in the stopband, rendering them suitable for applications with a smooth transition between the passband and stopband. However, they could fail to reduce the stopband frequencies enough, resulting in remaining noise in the filtered signal.

The Chebyshev filter is another filter that is widely utilized in the processing of EEG signals. It is meant to provide a steeper roll-off than the Butterworth filter but with greater ripple in the passband. The transfer function of the Chebyshev filter is described by,

$$H(s) = \frac{1}{1 + \epsilon^2 C_n(\frac{s}{w_c})^{2n}}$$

where is the ripple factor, Cn is the nth order Chebyshev polynomial, and the other variables are given the same meaning as the Butterworth filter. The frequency response of the The Chebyshev filter is characterized by,

$$H(f) = \frac{1}{1 + \epsilon^2 C_n(\frac{f}{w_c})^{2n}}$$

where f is the frequency parameter. (Chebyshev 1885). Chebyshev filters roll off more rapidly at the stopband than Butterworth filters, and that is advantageous for some applications. which require a larger attenuation in the stopband. They do possess ripples in the passband, which can influence the precision of the filtered signal.

The Elliptic filter is a filter that incorporates the best characteristics of both the Chebyshev and Butterworth filters. It has a steeper drop-off than the Butterworth filter and a flatter passband than the Chebyshev filter. The transfer function of the Elliptic filter is provided by,

$$H(s) = \frac{1}{1 + \epsilon C_n(\frac{s}{w_c})^{2n}}$$

where is the largest ripple in the passband, Cn is the nth order elliptic polynomial, s is the complex frequency variable, c is the cut-off frequency, and n is the order. of the filter. The frequency response of Elliptic filter can be expressed as below.

$$H(f) = \frac{1}{1 + \epsilon C_n(\frac{f}{w_c})^{2n}}$$

where f is the variable frequency. (Harris 1978). Elliptic filters have the steepest descent in the three filters' stopband and a smooth passband, so they are suitable for Applications that call for a steep cut in the stopband and a shallow slope. response in the passband. But they have ripples in both the passband and stopband, which can change the precision of the filtered signal.

These equations can be used in several EEG signal processing software such as EEGLAB (Delorme and Makeig 2004), MATLAB (Mathworks, Inc.), and Python-based packages such as MNE (Gramfort et al. 2013) and PyEEG (Python EEG signal processing toolbox) (https://github.com/forrestbao/pyeeg). These tools provide users with a set of filtering alternatives for EEG signals, such as the filter order, filter type, and cut-off frequencies. Typically, the selection of the filter is based on the individual characteristics of the EEG signals and the noise to be eliminated, including the range of interest of frequency and the level of attenuation in the stopband desired. Evaluating the performance of varying filters based on factors such as signal-to-noise ratio (SNR) and mean squared Error (MSE) assists in determining the optimum filter for a particular application.

- **Electrical noise removal**

Many techniques have been suggested and tried to minimize the Electrical line noise. The most common are filtering and regression methods.

Filtering is the most common technique used for noise reduction. Some of the filters that are widely used by individuals include notch filters, bandpass filters, and adaptive filters. Notch filters are very effective at removing narrow-band interference, especially the common 50 Hz or 60 Hz electrical noise from power lines.

Bandpass filters help to target a specific range of frequencies by removing frequencies outside of the desired range. Adaptive filters are special because they can adapt to remove different frequency components, making them ideal for real-time applications. A study conducted by Güler, Unal, and Akin (2018) examined these filtering methods to ascertain how efficient they were at removing electrical noise from EEG signals. They found that adaptive filters performed better, both at reducing noise as well as keeping the original EEG signal clear.

Apart from filtering, regression methods provide yet another effective way of reducing noise. These methods include linear regression, least mean square (LMS) adaptive filtering, and principal component analysis (PCA) (Widrow and Stearns, 1985; Pearson, 1901). These methods rely on the assumption that electrical noise can be separated from the brain signal and hence can be identified and removed. LMS adaptive filtering has also worked well in removing noise from EEG signals. This can be seen in studies by Li et al. (2020) and Khan, Khalid, and Javaid (2021). This method calculates the power spectrum of the noise and subtracts it from the noisy EEG signal. PCA has also worked well by breaking down the EEG data and separating noise from genuine brain activity (Li et al., 2020; Zhang et al., 2021).

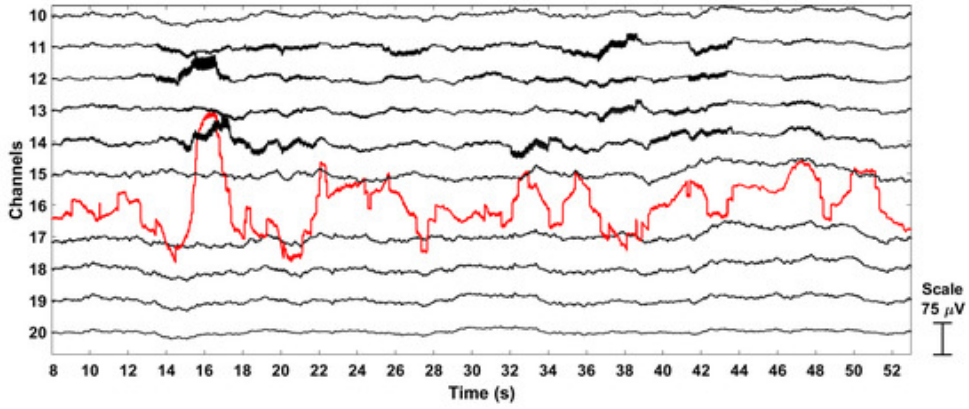
Each denoising method has its own strengths based on the nature of the noise. Notch filters are ideal for fixed frequency interference, e.g., 50/60 Hz, but might not deal well with wider or changing noise. Bandpass filters improve signals by eliminating frequencies outside of the brainwave range of interest, but some noise can remain. Adaptive filters are very adaptable, hence perfect for variable and complicated noise conditions. Regression-based techniques, such as LMS and PCA, perform well when the noise is dissimilar to the EEG signal. Individuals typically select these techniques when maintaining the quality of the signal high with most types of noise is crucial. The optimal selection is based on the type of noise and the objectives for preparing the EEG analysis.

- **Bad channels Detection and Removal**

  EEG signals can be influenced by different artifacts that can hide real brain activity and decrease the performance of models recognizing emotions. A common problem is the presence of "bad" channels, which can appear as a result of such problems as electrodes moving, not getting in good contact with the scalp, or other sources of external noise.

  Figure 2.9 is a sample simulated EEG trace. Channel 16 has been highlighted in red to signify that it is troubled by peculiar artifacts. It is thus unreliable for analysis. The remaining channels, which appear in black, are normal and are not affected by these problems.

Figure 2.9: Bad EEG Channel



Several approaches exist for identifying poor EEG channels. Perhaps the most prevalent is for skilled technicians to manually, visually inspect the signals to determine problems, including very high amplitudes or flat, unresponsive signals. Although effective, this approach can be time-consuming, based on subjective decisions, and have inconsistencies in results.

To avoid these limitations, automatic approaches have been devised. One is based on kurtosis (O'Reilly, Nielsen, and Hansen, 2007), in which channels with very high kurtosis values are identified as outliers. Correlation-based detection (Nolan, Whelan, and Reilly, 2010) and variance-based analysis (Viola et al., 2009) are also popular approaches that examine how each channel stands in relation to others and look for abnormal variability, respectively.

When the faulty channels are identified, they are usually removed or corrected before EEG analysis is continued. One of the usual techniques for correcting them is interpolation, where the data in the faulty channel is reconstructed using data from the surrounding electrodes. There are different techniques of interpolation, ranging from the simple linear interpolation to more complex techniques like spherical spline interpolation and multi-sphere head modeling (Fabien Perrin et al., 2011; Kayser and Tenke, 2006). The selection of an interpolation technique may have a significant impact on the final quality of the EEG signal.

Aside from interpolation, other removal strategies exist. Robust averaging (Pernet, Wilcox, and Rousselet, 2011), for example, involves averaging EEG signals across multiple trials but excluding data from bad channelsespecially useful when bad channels appear intermittently throughout trials. Another approach is threshold-based rejection, where channels with signal-to-noise ratios falling below a certain threshold are excluded (Mognon et al., 2011).

Bad channel detection and correction is one of the key steps in preprocessing EEG data to achieve appropriate emotion recognition. Several methods may be employed to detect and correct them, and all methods have advantages and disadvantages. Programs like EEGLAB (Delorme and Makeig, 2004), FieldTrip (Oostenveld et al., 2011), and MNE-Python (Gramfort et al., 2013) offer great platforms for researchers to remove bad channels effectively.

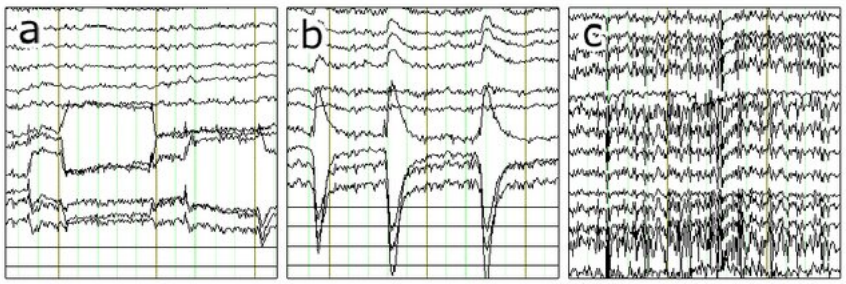
- **Artifcat Detection and Removal**

  An important function of preparing EEG data is detecting and eliminating undesired noise that reduces the quality of the signal as well as the reliability of results. It is extremely critical to detect and eliminate artifacts from emotion recognition applications that employ EEG.

  EEG signals may be influenced by various kinds of artifacts, such as eye movements, muscle contractions, and electrode movement. Eye blinks occur frequently and alter the frontal and temporal regions of the EEG. Muscle artifacts occur when the muscles move or contract, such as on the facial regions or the neck, and interfere with the adjacent electrodes. Electrode movement artifacts occur when electrodes move, resulting in indistinct and unsteady recordings. Since these artifacts have a significant influence on analysis, they must be identified and eliminated effectively.

  (Plass-Oude Bos 2012) did an analysis on identifying physiological artifacts in EEG. In figure 2.10, the first image(a) is eye movement, second image(b) is eye blink, and final image(c) indicates muscle tension.

Figure 2.10: Several types of Artifacts in EEG



  There are numerous methodologies for the identification and removal of undesired anomalies from EEG recordings. One of them is visual inspection based on examining the EEG traces for detecting anomalies. This method is effective but time-con.

  Another stronger method that operates independently is Independent Component Analysis (ICA). It separates EEG signals into independent components. We may identify and eliminate the components associated with artifacts by examining their location and frequency characteristics (Makeig et al., 2004).

  Another useful technique is wavelet analysis. It decomposes EEG signals into frequency bands, eliminating high-frequency noise or brief artifact (Cohen, 2014). It is particularly effective in eliminating artifacts with definite frequency patterns from the brain activity we are interested in. There are numerous special software programs that give end-to-end solutions for artifact removal. Some of them are EEGLAB (Delorme Makeig, 2004), FieldTrip (Oostenveld et al., 2011), and BrainVision Analyzer (GmbH, 2021). All of these software tools have sophisticated algorithms that assist with operations such as ICA, filtering, and automatic removal of artifacts.

  EEGLAB is a widely used free software tool for resolving many of the problems presented by data. One of its useful features is the extended Infomax ICA algorithm. The algorithm decomposes the EEG signal into independent components using

mathematical methods (Bell Sejnowski, 1995). The EEG signal appears as a combination of several sources, and the algorithm attempts to "unmix" the signals so that they are not dependent on one another by eliminating mutual information.

The Extended Infomax method employs a measure called maximum entropy in an effort to enhance statistical independence while preserving the essential time and frequency characteristics of the signal. For this reason, it is possible for it to identify and eliminate eye blinks and muscular movements because they possess unique frequency patterns.

Removal of artifacts is the most crucial initial step when doing EEG analysis, particularly for identifying emotion. Procedures such as ICA and wavelet de-composition, as well as useful software such as EEGLAB and BrainVision Analyzer, easily remove the EEG artifacts. One of these techniques, the extended Infomax ICA algorithm, is a useful method of decomposing the elements and eliminating various types of non-neural noise, that enhances the quality of the signal for subsequent analysis.

- **Re-referencing**

  Re-referencing is a fundamental pre-processing operation of EEG data that seeks to normalize the captured signals by converting them into a common reference point (Fabienne Perrin et al., 1989). Since EEG captures the voltage difference between scalp electrodes and a reference electrode, the initial reference selection may significantly influence the captured signal (Yao, 2001). Therefore, re-referencing is important to remove the bias introduced by the reference electrode and to make signal interpretation consistent (Sara M. Alarcão Maria J. Fonseca, 2017).

  Though widely used reference points such as earlobes, nose, or the average of all electrodes are widely used, they may impact the signal differently in some cases and result in varying results (Zhu et al., 2017). Thus, re-referencing helps prevent such differences and yields better data analysis.

  Two of the most common procedures to re-reference EEG recordings are the Average Reference (AR) and the Reference Electrode Standardization Technique (REST) (Zhu et al., 2017).

  - **The AR method** is to calculate the mean of all the electrode potentials and use this mean as a new reference. Simple as it is, this will introduce a shift in the global signal, and this might make subsequent analysis complicated in terms of accuracy (Fabienne Perrin et al., 1989).
  - **The REST method** offers a better solution. REST considers the physical layout of the electrodes and resorts to a simulated head model that generates reference-free signals. Despite being more complex than AR, REST has registered higher consistency and reliability, particularly for emotion estimation tasks (Zhu et al., 2017).

  To wrap up, re-referencing is necessary for enhancing the consistency and accuracy of EEG data towards the goal of emotion recognition. The method needs to be carefully selected, where REST is heavily employed because of its ability to give reference-independent results (Sara M. Alarcão Maria J. Fonseca, 2017). Packages

such as EEGLAB (Delorme Makeig, 2004), BrainVision Analyzer (GmbH, 2021), and FieldTrip (Oostenveld et al., 2011) include built-in facilities for using such re-referencing techniques to the best possible advantage.

2.4.3 Feature Extraction

Feature extraction is the conversion of meaningless data into meaningful features that reflect vital information in the data. Feature extraction in the identification of emotions from EEG tries to obtain features that reflect brain activity associated with emotion. Features are utilized as an input for the machine learning model in identifying the emotional state of the individual. Raw EEG recordings are cluttered and complicated, and detecting emotion from them is difficult. Feature extraction simplifies EEG signals and makes them easier to interpret. It also identifies significant content in EEG signals that aids in emotion identification. Our extracted features have the ability to indicate how the brain relates with emotion, making understanding emotion regulation possible. There are several ways to extract key features from the data in several ways for EEG-based emotion recognition systems. The most popular techniques are time-domain, frequency-domain, and time-frequency analysis.

- **Time-Domain Feature Extraction Method**

  Time-domain features are a basic part of EEG signal processing. Since most EEG recording instruments are recording signals in the time domain, an inspection of these unprocessed signals reveals the dynamics of neural activity over time. Features usually focus on amplitude, length, and waveform shape. The central objective of time-domain analysis is ultimately that of reducing the very high dimensionality of EEG signals with minimal loss of important information, making subsequent operations like emotion classification more efficient.

  Time-domain features play an important part in improving the efficiency of EEG processing, thus improving the performance of emotion recognition algorithms. Time-domain features can either be utilized individually or coupled with other approaches like frequency-domain and time-frequency analysis of the EEG in order to provide more accurate results. Some of the most widely used time-domain statistical features include mean, variance, standard deviation, skewness, and kurtosis. The above measures capture the central tendency, variability, and shape of the distribution of the signal effectively. Other features like root mean square (RMS), peak-to-peak amplitude, zero crossing (ZC), mean amplitude (MA), integral of absolute value, and autocorrelation provide further insight into the energy, fluctuations, and rhythm of the signal.

  More sophisticated time-domain features include histogram analysis, wherein the EEG values' distribution is illustrated, thus making patterns or abnormalities easier to identify. Kurtosis quantifies the pointed nature of the peak of the signal when compared with a normal distribution, while skewness reveals any asymmetry of the shape of the signal. The fractal dimension, called the Hurst exponent, indicates the long-memory or intricateness of a time series. Entropy is one other effective measure that captures the randomness and irregularity of EEG fluctuationsfeatures which are of particular relevance when examining emotional and cognitive function-

ing. Overall, features in the time domain are necessary for filtering through raw EEG data while maintaining the underlying patterns of brain activity. Not only do they reduce the dimensionality and noise of the data, but they also provide useful information that aids the construction of efficient emotion recognition systems based on EEG.

- **Frequency-Domain Feature Extraction Method**

  Frequency-domain feature extraction involves altering EEG signals from time-based into frequency-based data in order to view their frequency features. This transformation enables researchers to view the power spectral density (PSD), indicating the power of signals spread over the frequency bands. Since some of the frequency bands of EEG signals correlate with thoughts and feelings, observing them provides critical information for emotion identification.

  For instance, alpha and beta bands are associated with pleasant feelings, whereas theta and delta bands are associated with unpleasant feelings. By isolating these bands and examining them, scientists are able to discover indications of what a person feels. This makes frequency-domain features extremely useful for emotion-detecting systems. A popular method of calculating PSD is with the Fast Fourier Transform. It de-composes the EEG signal into frequency components. We have the option of applying measures such as mean, variance, and skew on the transformed signal. Wavelet transforms are utilized frequently because they both show time and frequency simultaneously. It becomes easy for us to identify particular bands associated with emotional activity. The PSD is computed by the autocorrelation function of EEG signal using Welchs method. The information about the sequence is as follows:

$$X_i(n) = x(n + iD)$$

  here n = 0, 1, 2. . . ., M-1, and i = 0, 1, 2. . . .. L-1. If Xi (n) is the sequence, iD will be the first point, and L shows the length of 2M, which is a segment of information. The output is presented as,

$$p_{xx}^{\approx(t)}(f) = \frac{1}{MU}\left|\sum_{n=0}^{M-1} x^i(n)w(n)e^{-j2\pi fn}\right|^2$$

  In the above window function, U is the regularization feature of the power and it is denoted by,

$$U = \frac{1}{M}\sum_{n=0}^{M-1} w^2(n)$$

  Here, w (n) is window function that describes Welchs power spectral as,

$$p_{xx}^{(W)} = \frac{1}{L}\sum_{t=0}^{L-1} p_{xx}^{\approx(t)}(f)$$

PSD isn't the only key attribute in the frequency-domain. Other key attributes include higher-order spectra (HOS), higher-order crossings (HOC), differential entropy (DE), and the logarithmic energy spectrum. Differential entropy(DE) addresses the intricacies of continuous EEG signals, and within specific frequency bands, DE will equal the logarithmic energy. Studies have indicated that regions of EEG signals that have been filtered using band-pass filters typically exhibit a Gaussian distribution, making DE a useful method for measuring intricacies in these regions.

$$DE == -\int_X f(x)log(f(x))dx$$

The above expression measures the differential entropy, where X is a random variable and f(x) is the probability density function of the experimental studies.

- **Time-Frequency Domain Feature Extraction Method**

Time-frequency domain analysis approach merges information from both time domain and frequency domain and has local analysis capability in time-frequency domain simultaneously.

The original signal time-domain information will not be lost in frequency domain analyses of the EEG signals, and in analyses, it is also possible to preserve greater resolution.

Short-time Fourier transform (STFT) - is a highly popular time-domain feature extraction technique. It comprises breaking up the EEG signal into short intervals and performing a Fourier transform on each one segment. STFT provides time-frequency characterization of the signal and can be used to It extracts features such as spectral power, spectral entropy, and spectral centroid. Non-stationary process is considered to be a sum of an aggregation of short-time stationary signals. Its calculation formula can be expressed as:

$$X(t,f) = \int_{-\infty}^{+\infty} x(u)w(u-t)e^{-j2\pi fu}du$$

Where w(u-t) is the short time window function. The fixed size and shape of the window function make it incompatible with high-frequency temporal subdivision and low-frequency subdivision needs.

The wavelet transform (WT) - inherits the STFT's capacity to do local analysis. The Fourier Transform breaks down a signal into its sinusoidal components. But, it lacks time localization. This means, if a brief but significant event occurs in a signal, the Fourier Transform might not capture it effectively. Enter wavelets. With their localized nature, wavelets can capture both frequency and time information. This dual nature makes them especially suited for non-stationary signals, where the signal's properties change over time. It decomposes the EEG signal into a set of wavelet coefficients at different scales and positions in time. WT can be used to extract features such as wavelet entropy, wavelet energy, and wavelet variance.

The Discrete Wavelet Transform (DWT) is one of the commonly used methods for time-frequency-domain feature extraction in EEG-based emotion recognition. The signal is reconstructed using DWT. The DWT decomposition produces two sets of functions on both sides: scaling coefficients (xk+1(n)) and wavelet coefficients (yk+1(n)) (Pooja, Pahuja & Veer 2022). The following equation illustrates the use of two coefficients.

$$X^{k=1}(n) = \sum_{i=1}^{2n} h(2n-i)X^k(n)$$

$$Y^{k=1}(n) = \sum_{i=1}^{2n} h(2n-i)X^k(n)$$

Where k denotes the scaling coefficient. It offers a multi-resolution description of the signal that is useful in representing low- and high-frequency components at various scales. DWT is generally applied in EEG systems for extracting features like delta (0.54 Hz), theta (48 Hz), alpha (813 Hz), beta (1330 Hz), and gamma (>30 Hz) rhythms. An advantage of DWT is that it is computation efficient, which is appropriate for use in real-world systems and in embedded systems. Apart from that, signal feature compact representations help in dimensionality reduction that could assist machine learning models to avoid overfitting. Statistical features like mean, variance, skewness, and entropy of DWT coefficients could also prove helpful in representing emotional state and cognitive activity.

Continuous Wavelet Transform(CWT) provides a clearer additional version of the signal by computing wavelet coefficients at every possible place with all scales. It serves to examine the signal closely, which is why CWT is so useful for displaying time-frequency patterns along with abrupt changes in EEG signals.

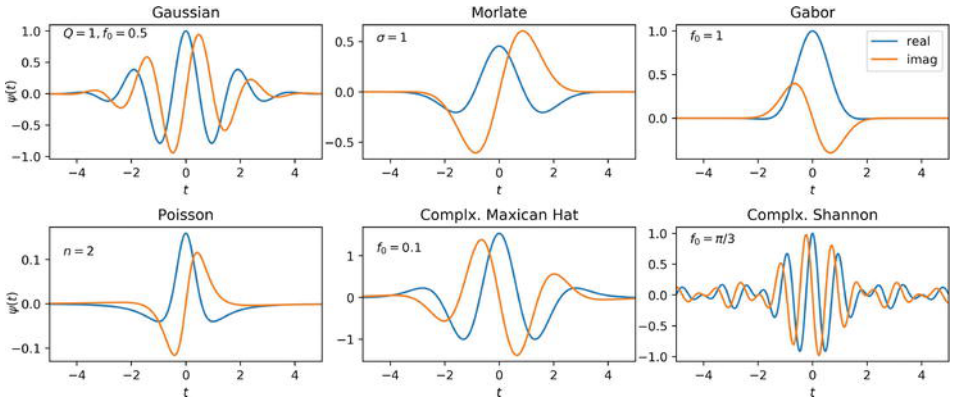
In deep learning systems, we can use CWT-based techniques, particularly with 2D images referred to as scalograms. They can be fed to CNNs to identify emotion or illnesses. Although it needs more computational demands compared to DWT, it enables better time-frequency information and is extremely useful in situations in which small changes in brain activity need to be noticed.

$$CWT_x(a,b) = \int_{-\infty}^{+\infty} x(t).\frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right)dt$$

Where x(t) is the input signal, ψ is the mother wavelet function, a is the scaler function that give information about the frequency component and b is the translator factor which gives information on the time localization. CWT uses mother wavelets as its fundamental functions for signal analysis and representation. Using scaling and translating operations, a mother wavelet which is a finite-energy function with zero meanis utilized to create a family of wavelets. The properties of the signal being examined and the required time-frequency resolution determine which mother wavelet is used. To extract time-frequency information, the CWT compares the signal with translated and scaled version of the mother wavelet. Following are some of the Mother Wavelet used in CWT(Figure 2.12).

- **Gaussian Wavelet:** A time-domain wavelet, it is derived from a Gaussian function centered at t0 and modulated by a complex exponential function with frequency f0.
- **Gabor Wavelet:** Gabor wavelet is perhaps the most widely used function for various applications. It is essentially the same as Gaussian wavelet function, with simplified equations.
- **Morlet Wavelet:** Morlet is considered very similar to Gabor wavelet and Gabor filters. The oscillation of Morlet wavelet is controlled by . A higher value of results in higher oscillation.
- **Poisson Wavelet:** Poisson wavelet is defined by positive integers (n), unlike other, and associated with Poisson probability distribution.
- **Complex Mexican hat wavelet:** Complex Mexican hat wavelet is derived from the conventional Mexican hat wavelet. It is a low-oscillation wavelet which is modulated by a complex exponential function with frequency f0.
- **Complex Shannon wavelet:** Complex Shannon wavelet is the most simplified wavelet function, exploiting Sinc function by modulating with sinusoidal, which results in an ideal bandpass filter. Real Shannon wavelet is modulated by only a cos function.

Figure 2.11: Mother Wavelets used in CWT



2.4.4 Feature Selection

Feature selection is required for emotion recognition based on EEG due to several reasons. First, EEG data is inherently high-dimensional with hundreds if not thousands of features recorded from multiple electrodes. It retains irrelevant or redundant features that can lead to overfitting and reduce the ability of machine learning models to generalize. Feature selection can also reduce computational costs for the analysis and enhance the system's scalability. Last but not least, feature selection can enhance interpretability of the outcome by determining the most significant features that are relevant to the brain processes that deal with emotion. There are various kinds of feature selection techniques that are applicable to EEG-based emotion recognition depending on the particular study question and type of data. Most often, Feature selection can be broadly classified under two main categories,

- Filtration methods
- Wrapper Methods

Filtering techniques assess each of these features regardless of the classification algorithm and choose the highest-ranked features according to some predetermined criterion. The the most commonly used filter methods are mutual information (Ross, Jaroszewski, and Schmidhuber Styczynski 2014), correlation based feature selection(H. Liu 2018), and (Agresti Finlay 2009). These methods estimate the relevance of each feature with for the target variable (i.e., emotional state) and select the top-ranked features dependent upon a threshold or a specified number. On the contrary, wrapper methods make feature selection an intrinsic process of classification algorithm and rank them based on their performance under a cross-validation procedure. Wrapper methods are computationally more expensive but can offer better performance than filter methods by incorporating feature selection process into the classification algorithm. Most common wrapper approaches in EEG-based emotion recognition are genetic algorithms (Goldberg 1989), particle swarm optimization(Kennedy and Eberhart 1995), and Sequential Forward/Backward Selection(Kohavi and John 1997).

2.4.5 Classification

Emotion Classification from an EEG signal refers to determining what people feel based on their brain activity patterns. We use the EEG signal, and we have to assign to it an appropriate emotional label according to its features. Various approaches, ranging from simple machine learning to sophisticated deep learning models, are applied to accomplish this task.

- **Traditional Machine Learning Approaches**

  Traditional learning techniques such as Support Vector Machine (SVM), k-Nearest Neighbours (k-NN), and Naive Bayes (NB) are generally applied to identify emotions from EEG.

  - **SVM** determines an optimal hyperplane to establish maximum distance between various emotion classes.
  - **KNN** is a straightforward approach that assigns a label to new data based on looking at which label is most commonly found among its k nearest points.
  - **Naive Bayes** applies probability to predict the likelihood of a data point belonging to some class, given the features being independent.

- **Deep Learning Techniques**

  Currently, deep learning algorithms are gaining popularity due to deep models' ability to learn complicated patterns from raw EEG data independently.

  - **Convolutional Neural Networks (CNNs)** excel at detecting patterns and shapes with the use of specialist filters.
  - **RNNs, such as LSTMs**, are well suited for handling sequential EEG signals since these can extract time-related information from the signal.

- **Datasets Used for Understanding Emotions**

  In order to validate these classification techniques, a number of publicly available EEG datasets are tested. Some standardly used benchmarks include DEAP, SEED, and DREAMER.

  - **DEAP Dataset:** This dataset was developed by Koelstra et al. in 2012. It contains EEG recordings along with other recordings from 32 individuals viewing 40 one-minute music videos. 32 channels of EEG, EOG, and EMG signals are present too. These can be used to investigate emotion based on diverse aspects. Each video recording begins with 3 seconds of baseline.
  - **SEED Dataset:** SEED dataset was captured by the BCMI lab with 15 individuals. They presented every individual with fifteen emotional movie clips three times. They recorded with 62 electrodes at 200 Hz. This dataset is suitable for exploring stable emotional patterns over multiple sessions.
  - **DREAMER Dataset:** Katsigiannis and Ramzan developed this dataset in 2017. 23 individuals were recorded with EEG while viewing 18 emotional movie clips. Each individual rated their emotion for valence, arousal, and dominance. 14 channels of EEG were recorded with a 128 Hz sample frequency.

2.5 Emotion Recognition(ER) in Song Vocals

In the domain of ER in music or on song vocal most of the researchers have used methods based on Physiological responses. Lets go through some literatures and their limitations.

In 2011 (Pell & Kotz 2011) introduced a method, which utilizes the auditory gating paradigm to study the temporal dynamics of vocal emotion recognition, which provides significant insights into how listeners recognize emotions from speech as it unfolds. By segmenting pseudo-utterances into increasing syllable durations and analyzing recognition patterns across six emotions (anger, disgust, fear, sadness, happiness, and neutral), the study effectively maps the progression of recognition accuracy and confidence. It highlights that fear, sadness, and neutral expressions are recognized faster and with greater accuracy at early stages, whereas happiness and disgust take longer and show more variability in recognition. The integration of acoustic measures at the "identification point" further illuminates the specific features, such as pitch and speech rate, that characterize each emotion's unique recognition trajectory.

So many limitations can be found in this model. First, gating by syllable duration instead of fixed time intervals might introduce variability in temporal estimates due to differences in emotion-specific syllable lengths. This could skew recognition times, particularly for emotions like sadness or disgust with longer syllables. Second, the reliance on pseudo-utterances, while controlling for semantic bias, may not fully replicate the complexities of natural speech, limiting the generalizability of findings to real-world communication. Finally, forced-choice tasks, while useful for categorization, may not capture the nuanced and dynamic nature of emotional perception in less constrained listening conditions. These factors underscore the need for further methodological refinements and complementary approaches to build on the study's findings.

EmoMucs(Berardinis, Cangelosi & Coutinho 2020), a novel computational model designed to enhance Music Emotion Recognition (MER) by integrating music source separation techniques. The model decomposes a music track into separate sources (vocals, drums, bass, etc.) and processes each component through specialized sub-models to predict emotions based on valence and arousal. By combining these predictions using various fusion strategies, EmoMucs provides improved interpretability and accuracy in predicting emotional responses to music. Tested on the PMEmo dataset, EmoMucs outperformed existing deep learning models in valence prediction while offering comparable results for arousal. A significant advantage of EmoMucs is its modular design, which facilitates tailored analysis of the emotional contributions of individual musical elements. But EmoMucs model(Berardinis et al. 2020), while innovative, faces challenges related to its dependence on the pre-trained Demucs system for source separation, which might introduce errors if the separation quality is poor. The modular design, though flexible, increases the computational complexity and training time compared to simpler models. Additionally, the model's performance improvements are limited for arousal prediction, where it achieves results comparable to baseline methods. The lack of a systematic evaluation of the interpretability aspect in real-world settings and reliance on static annotations further limits its broader applicability to dynamic, real-time emotional analysis.

CFIA-Net(Hu, Yang, Huang & He 2024), introduces a Cross-modal Features Interaction-and-Aggregation Network with a self-consistency training strategy for speech emotion recognition (SER). It leverages audio and textual features extracted via pre-trained models like emotion2vec and BERT. The CFIA module ensures effective integration of multimodal data through adaptive interaction and aggregation, while the self-consistency training supervises shallower layers with deeper ones to enhance feature learning without increasing model complexity. Experimental results on the IEMOCAP dataset demonstrate state-of-the-art performance with weighted and unweighted accuracies of 83.37% and 83.67%, respectively, outperforming existing bimodal SER methods.

However, this method also has some limitations. It heavily relies on the quality of pre-trained models and may face challenges when applied to datasets with different distributions or noise characteristics. The computational overhead of the CFIA(Hu et al. 2024) modules and self-consistency strategy might limit its scalability to real-time applications or low-resource environments. Additionally, while effective for bimodal audio-text integration, its adaptability to other modalities like video or physiological signals remains unexplored. Further research could explore these aspects to enhance the robustness and generalizability of CFIA-Net.

So we can come to the conclusion that existing models for vocal emotion recognition, such as those utilizing acoustic features or behavioral paradigms, often face limitations in capturing the nuanced, dynamic, and neural underpinnings of emotional perception. These approaches primarily rely on explicit responses or acoustic pattern analysis, which may not fully account for the rapid and subconscious nature of emotional processing in real-world scenarios. In contrast, using EEG to assess emotional responses to vocal stimuli offers significant advantages. EEG provides high temporal resolution, enabling researchers to capture the brain's immediate reactions to emotional prosody as it unfolds. This approach goes beyond surface-level acoustic analysis, allowing for the exploration of neural markers associated with discrete emotions, such as event-related potentials (ERPs). By directly tapping into the neurocognitive processes underlying emotion recog-

nition, EEG can provide a more comprehensive understanding of how the brain processes vocal emotions in real-time, overcoming the limitations of models that focus solely on external behavioral or acoustic cues.

2.6 Emotion to Melody Generation

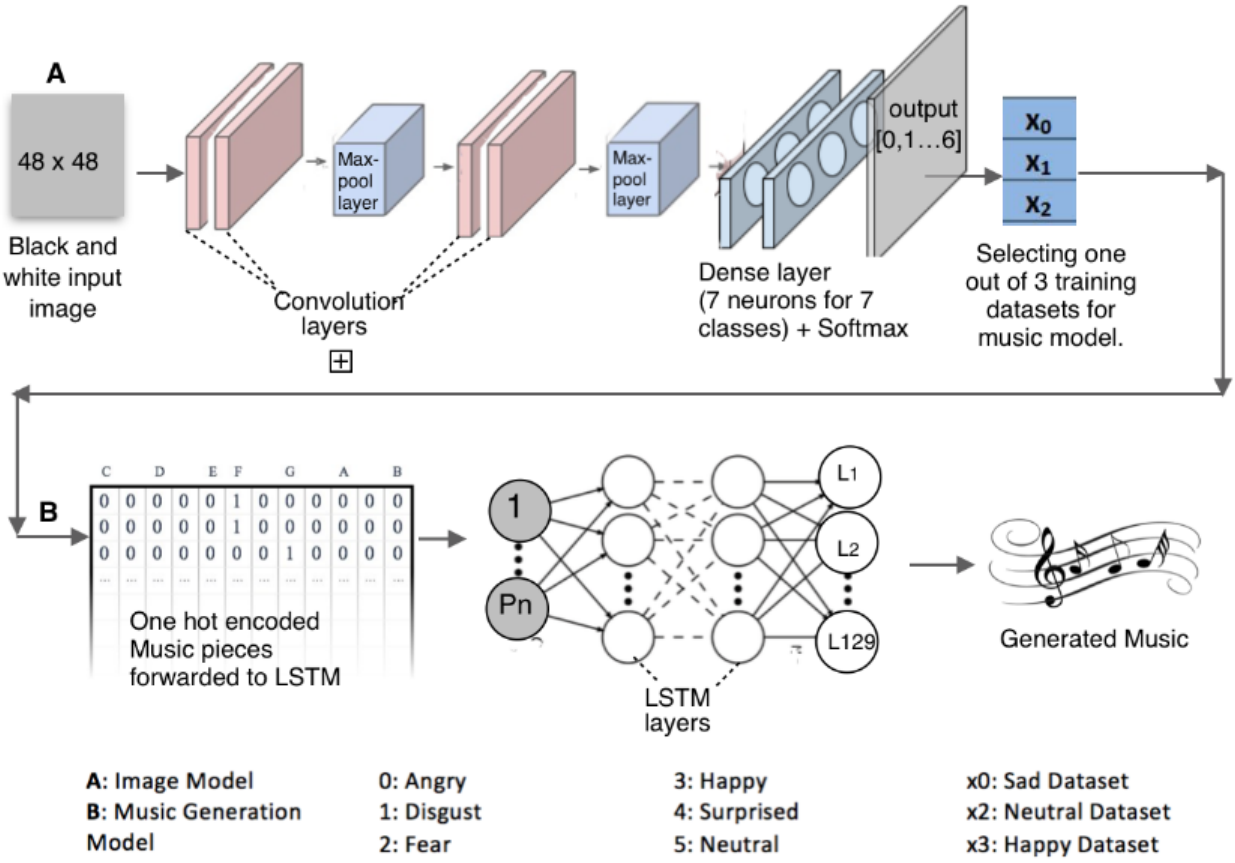
In this section, I'll be looking into some state of the art standard models for Emotion based music generation models and their limitations. The EmotionBox model (Zatorre, Mori, Sang & Wang 2022) aims to advance emotional music generation by leveraging music psychology principles and deep neural networks. Unlike previous label-based methods that depend on extensive emotion-labeled datasets, EmotionBox uses intrinsic musical featuresnote density and pitch histogramto control arousal and valence, respectively. These features are mapped to specific emotions using the Russell emotion model, enabling the generation of music in four emotional categories: happy, sad, tensional, and peaceful. The use of a classical piano MIDI dataset simplifies preprocessing and eliminates the need for manual emotion labeling. Subjective listening tests demonstrate that EmotionBox performs comparably to traditional label-based systems and is particularly effective in generating low-arousal emotions like peaceful and sad, highlighting its potential for applications such as music therapy.

However, the model(Zatorre et al. 2022) has notable limitations. Valence representation remains challenging, as the reliance on mode (major or minor) inadequately captures its complexity. Additionally, the generated music lacks structural coherence, resembling improvisation rather than complete compositions. The homogeneity of the training dataset, limited to classical piano pieces, restricts the model's adaptability to diverse musical genres and instruments. While effective for low-arousal emotions, high-arousal emotions such as happy and tensional are less distinct, pointing to variability in performance. Future work should focus on enhancing valence features, incorporating long-term structural patterns, diversifying training datasets, and balancing subjective evaluations with objective metrics for a more comprehensive assessment.

SentiMozart (Madhok, Goel & Garg 2018) framework highlights its dual-purpose structure: capturing human emotions via facial expression analysis and generating corresponding music. The system employs a Convolutional Neural Network (CNN) for sentiment classification into seven categories, which are further grouped into three main classes (Happy, Sad, Neutral) for music generation. The music generation model uses a Doubly Stacked LSTM architecture, trained on a manually annotated dataset of MIDI files. The framework's performance is evaluated using the emotional Mean Opinion Score (MOS), revealing a strong correlation (0.93) between detected facial sentiments and the generated music, indicating its efficacy in sentiment-aligned music generation.

Also SentiMozart (Madhok et al. 2018) face several limitations. Firstly, the reliance on manually labeled MIDI files may introduce subjective bias and restrict scalability, as the dataset requires significant human effort to expand. Secondly, the CNN's sentiment classification, while achieving an accuracy of 75 percent, may struggle with more complex emotional nuances or diverse datasets. Additionally, the LSTM-based generation model might face challenges in capturing the intricate global and local musical patterns due to the inherent complexity of music composition. Lastly, the MIDI format, while efficient in size, produces inconsistent audio quality across systems, potentially reducing the end-user

Figure 2.12: SentiMozart Model Architecture



experience. These limitations suggest areas for improvement, such as incorporating larger, more diverse datasets and exploring advanced architectures to enhance both classification accuracy and musical creativity.

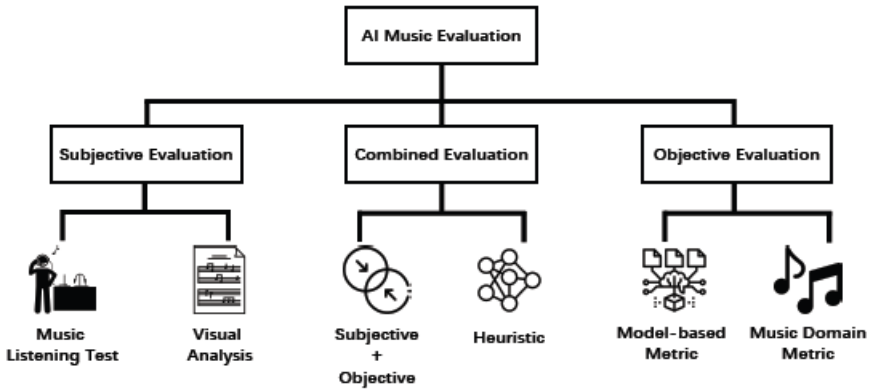
Finally want to mention about MusicGen(Copet, Kreuk, Gat, Remez, Kant, Synnaeve, Adi & Défossez 2023) model which is a single-stage transformer-based model designed for conditional music generation. Unlike traditional multi-stage models, MUSICGEN employs a single language model operating over compressed discrete music representations (tokens). It integrates text or melody conditioning, allowing users to generate music aligned with specific instructions or melodic features. The model leverages an autoregressive decoder and efficient token interleaving strategies to produce high-quality stereo and mono music at 32 kHz. Extensive evaluations reveal that MUSICGEN outperforms other baselines in text-to-music tasks, ensuring better control and adherence to the provided conditions. Its key advancements include simplifications in model architecture and new conditioning approaches, making it robust and versatile for music creators. So according to the literatures this model has been a state of the art model to get music generated according to condition on melody and prompt texts. So if tend to use this model by (Copet et al. 2023) for the proposed task we have to use emotions as inputs, the model can be conditioned on features that represent arousal and valence values derived from the desired emotional state. These values could guide the generation process by aligning them with corresponding melodic or harmonic structures. For instance, high arousal and positive

valence might translate into upbeat and energetic music, while low arousal and negative valence might result in slow and somber compositions. By mapping emotional inputs to chromagrams or textual descriptors that convey the intended mood, MUSICGEN can produce music that reflects specific emotional landscapes, thus enabling personalized and emotion-driven music creation.

2.7 AI Music Evaluation Methods

After delving into the two main topics in the literature, when discussing about the evaluation methods, almost all of the generative music studies evaluate their respective models under a Subjective or an Objective perspective. A survey on evaluation methods of AI-generated music models done by (Xiong et al. 2023) categorizes the evaluation criteria under three main categories as presented in Figure 2.13.

Figure 2.13: Evaluation Structure Proposed by (Xiong et al. 2023)



Most of the AMG have used music listening tests to evaluate their model one way or another. The researchers of the MusicVAE(Roberts, Engel, Raffel, Hawthorne & Eck 2018), gave participants two 30-second musical pieces, one from the produced piece and the other from the original, and asked them to rank which one they felt was more musical on a Likert scale. (Chu, Urtasun & Fidler 2016) surveyed 27 music professionals by providing them with several pairs of 30-second melody pieces and asked for a vote to decide which piece was better in the pair. Over 84% voted the piece generated by their model is better than the other generated models they have used to pair up against. Komposer (Dias & Fernando 2019) mentioned in his work, that user criticism should be acquired because computers and algorithms are incapable of evaluating music. They implemented a web-based inference tool to gather user feedback and examined it to determine the reliability and correctness of the outputs that were produced. Comments were gathered from both music professionals and amateurs.

The objective evaluation involves analyzing the generated melody using computation to produce metrics that can be measured of its quality. Meaurements of AMG qualities derived from the musical concept are considered as Music Metrics Evaluation. (Ji, Luo

& Yang 2020) categorized the metrics under Pitch, Harmoney, Rythm, and Style-related matrices. (Dias & Fernando 2019) introduced a novel algorithm called the Consistency Evaluation Algorithm. They reasoned that if ABC notation is used to train the model, the resulting output ought to adhere to the standard, teaching the model some fundamental principles of music theory. Specifically, they examined the trained model's ability to understand that for each melodic sector, the number of notes in the produced output should be the same. In other words, each sector of a 4-by-4 melody should have four notes. Using the suggested algorithm, they tallied the notes that were included in each sector and group, as well as the quantity of notes included in each component that had the same amount of notes. They then identified the group that appeared most frequently and used the equation to calculate consistency. c = no of sectors in the most frequent group/total no. of sectors

As you may clearly see current models typically struggle when attempting to match vocals and melodies on a more profound emotional level, particularly when the vocals are self-reflective. They tend to focus more on the technical details of the generated melodies more than the emotional response that the vocals convey. Because of this, it could be difficult for listeners to emotionally relate to a particular musical piece.

This study seeks to address this problem by developing a novel approach to melody generation that is based on the emotional responses of listeners interacting with song vocals. The emotional feedback of people listening to vocals will be use as the input for the melody generation model. Above approach promises to open up new avenues for music personalization. We can make customized, emotionally charged music that truly connects with each listener by comprehending and addressing the emotional landscape of song vocals.

Chapter 3

Methodology

3.1 EEG Emotion Recognition Process

3.1.1 Data Acquisition Experiment

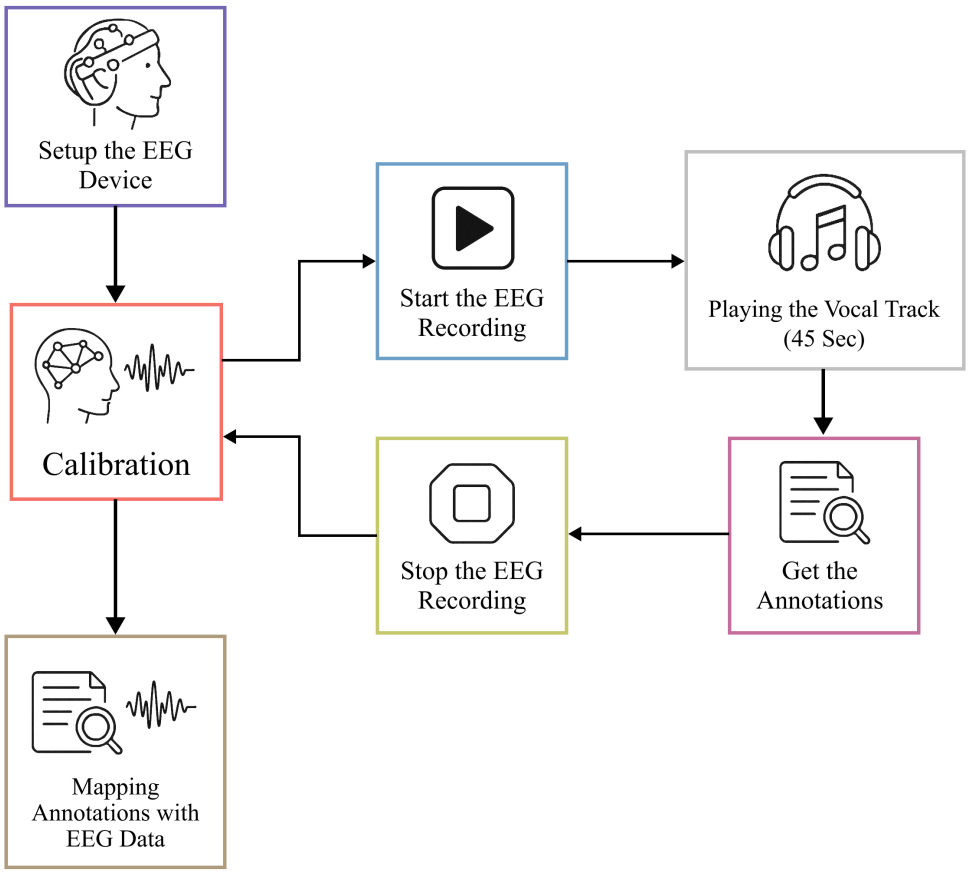
The EEG data acquisition process involved multiple steps using the Emotiv Pro software and EEG headset. The headset was first connected, calibrated, and assessed to ensure 100% contact and EEG signal quality. Participants were then instructed to close their eyes while listening to each song's vocals. Immediately after listening, they annotated their arousal and valence levels. This process was repeated for all songs, and the corresponding annotations were mapped to the EEG recordings before proceeding to data pre-processing.

Song and Subject Selection
A total of 104 songs were used in this study, comprising 65 Sinhala and 39 English tracks. The songs were carefully selected to represent the full spectrum of emotions within the arousal valence dimensional model. The dataset includes high arousal positive valence songs (e.g., energetic dance tracks), high arousal negative valence songs (e.g., intense or aggressive pieces), low arousal positive valence songs (e.g., calm and relaxing tunes), and low arousal negative valence songs (e.g., melancholic or sad ballads).

To determine the placement of each song within the arousal valence space, user comments from the corresponding YouTube videos were extracted and analyzed using a custom Python based script. These comments, which reflect the emotional reactions of listeners, were processed to evaluate sentiment polarity and identify emotion related keywords associated with arousal (e.g., energetic, chill, relaxing, intense) and valence (e.g., happy, sad, beautiful, painful). Each song was then assigned an estimated arousal valence position based on aggregated comment analysis. The study involved 33 participants, all university students aged between 20 and 26. Each participant was randomly assigned 10 song vocals from a pool of 104 tracks. Each song vocal lasted 45 seconds, and participants were instructed to listen attentively with their eyes closed. After each track, they provided self assessed arousal and valence ratings. The complete session for each participant, including EEG headset setup, calibration, and signal quality checks, took approximately one hour.

With each participant annotating 10 songs, each song was annotated by approximately

Figure 3.1: Overall Data Acquisition Experiment



three different individuals, ensuring redundancy and reliability in the emotional ratings. Prior to classification, both the EEG signal features and the arousal valence annotations were averaged across annotators for each song. This averaging helped reduce individual bias and variability, providing a more consistent representation of emotional response.

EEG Device and Annotation Model

The EEG data were recorded using a 32-channel Emotiv EPOC Flex headset(figure 3.2 & 3.3) and Emotiv Pro software. The headset was connected to the software either by USB or Bluetooth, and it was preferable to use USB since it provided a more robust and stable signal. EEG Gel was applied to the electrodes in order to ensure the electrodes came in contact with the scalp effectively. The sampling rate was configured at 128 Hz throughout the experiment.

For emotional labeling, we employed the Self Assessment Manikin (SAM) model(figure 3.4, which was derived from Russell's arousal valence theory. Individuals provided ratings for their arousal and valence feelings immediately following every EEG recording, enabling us to closely and continuously monitor their emotional responses.

Figure 3.5 show three participants carrying out the experiment.

Figure 3.2: Emotive EPOC Flex Headset Components

Figure 3.3: Emotive EPOC Flex Headset plugged in

3.1.2 Data Pre-processing

For the data pre-processing EEGLAB (Delorme and Makeig, 2004) WAS used, which was an extension for MATLAB Software to handle EEG Data. To enhance the quality and reliability of the EEG data, a comprehensive pre-processing pipeline was applied using EEGLAB. The pre-processing steps are summarized in Figure 3.6 and detailed below.

Initially, select the EEG channels and configure their respective scalp locations. Event markers were then identified to segment the EEG data appropriately.

The slow changes and high frequency noise were eliminated using Finite Impulse Response (FIR) bandpass filter spanning 0.5 to 50 Hz, leaving critical frequency components associated with thoughts and feelings in place. Line noise and other unwanted interference were eliminated by the use of the CleanLine algorithm.

Following line noise removal, re-referencing was performed to standardize the EEG signals and reduce reference related bias across channels. was then conducted to decompose the EEG signals into statistically independent components. The components were automatically labeled using the ICLabel plugin, which classifies components based on their source (e.g., brain, eye, muscle, line noise, channel noise, and others). Non neural components (e.g., those representing ocular, muscular, or noise artifacts) were subsequently flagged as artifcats and removed based on ICLabel confidence scores. Finally, any boundary events and discontinuities in the EEG recordings were removed to ensure clean and continuous data segments for further analysis.

This multi stage pre-processing approach significantly improved the signal to noise ratio,

Figure 3.4: Self Assessment Manikin(SAM

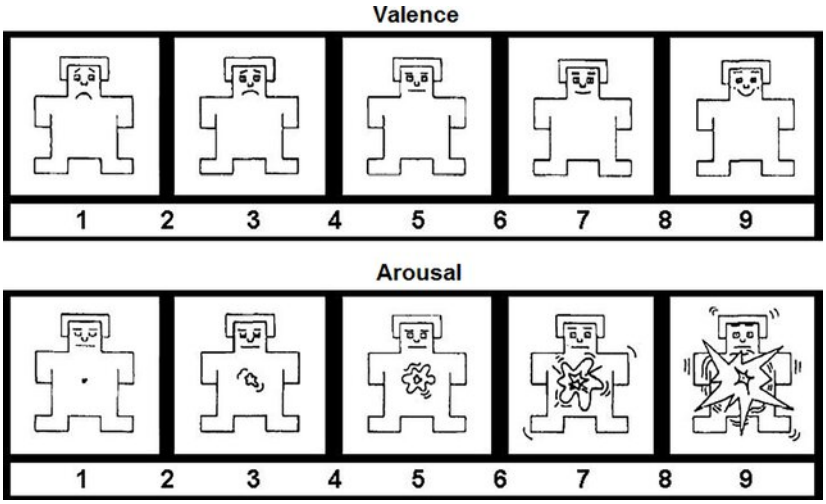


Figure 3.5: Snapshot of participants during data collecting experiments

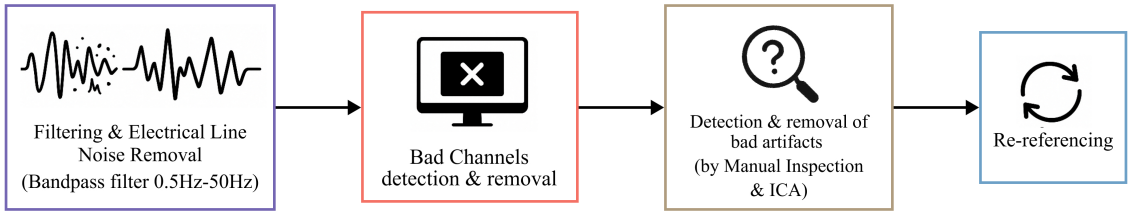
reduced the presence of artifacts, and ensured the data was clean, consistent, and ready for feature extraction and classification.

3.1.3 Feature Extraction

EEG signals are inherently non stationary and composed of multiple oscillatory components distributed across distinct frequency bands which are delta (0.5 - 4 Hz), theta (4 - 8 Hz), alpha (8 - 13 Hz), beta (13 - 30 Hz), and gamma (>= 30 Hz). These bands are functionally associated with different cognitive and emotional states; for instance, delta waves are linked to deep sleep and unconscious processing, while alpha activity is typically observed during relaxed wakefulness and visual attention. Analyzing the spectral power distribution across these bands offers critical insight into the participant's neural and emotional states during auditory stimulation.

To capture these dynamic characteristics, wavelet transforms were employed due to their ability to provide localized information in both time and frequency domains. In this study, the primary focus was on the Continuous Wavelet Transform (CWT) using the Morlet wavelet, owing to its excellent time frequency resolution and suitability for analyz-

Figure 3.6: Pre-processing Methodology



ing transient oscillatory patterns in EEG. The Morlet wavelet was particularly effective in highlighting subtle changes in low frequency bands, as observed in the resulting scalograms.

The transformation to morlet wavelet was done within a scale range of 1 to 64, which corresponds to different frequency resolutions. Each EEG channel was independently transformed, producing a detailed time frequency representation for the entire signal. Scalograms generated using CWT revealed significantly more nuanced temporal patterns compared to traditional spectrograms. Moreover, wavelet functions such as Poisson, Complex Shannon, and Complex Mexican Hat were also experimented with. While similarities were observed among the latter two, Morlet and Poisson wavelets offered enhanced resolution at lower frequencies, which are particularly relevant in emotional EEG studies.

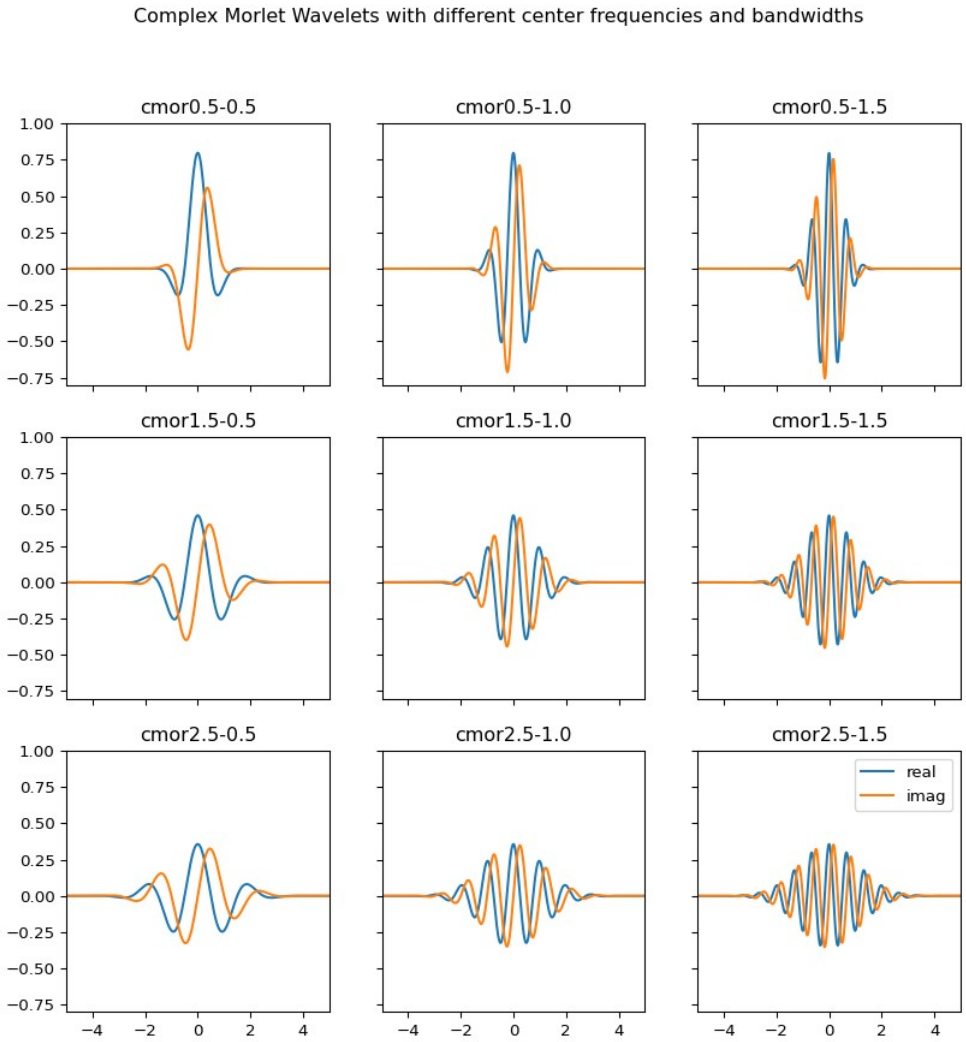
To benchmark the performance of Continuous Wavelet Transform (CWT), a comparative analysis was also conducted using Discrete Wavelet Transform (DWT). Unlike the Fourier Transform, which provides uniform frequency resolution, DWT adapts to the frequency characteristics of the signal by varying time and frequency resolution based on scale. This makes it especially effective for analyzing non stationary EEG data.

In the DWT approach, the EEG signal was decomposed into wavelet coefficients at different scales using mother wavelets tailored to the properties of the EEG signal. These coefficients represent the energy localized in specific frequency bands and time intervals. Band specific spectral powers corresponding to delta, theta, alpha, beta, and gamma bands were computed by summing the squared magnitudes of coefficients within each respective band. These features were then used for classification and correlation analysis. While DWT allowed for efficient extraction of frequency band specific features and was computationally less intensive, CWT provided a richer and more continuous view of EEG dynamics. Ultimately, this dual approach enabled both a robust benchmark comparison and a deeper understanding of how wavelet selection impacts feature representation in EEG based emotion recognition.

3.1.4 EEG Feature Classification

In the final step of using EEG to recognize emotions, the features are classified into categories like Happy, Sad, Angry, and Calm. But, this taxonomy makes it too simple for the complex range of human emotions. People often find it hard to tell apart emotions that are very similar. For example if someone feels sad and calm kind of a sense for some stimuli, he might get confused. It leads to ambiguity and reduced accuracy for the classification model. In order to resolve this problem, this study utilizes Russell's Arousal Valence Model, which represents emotions in two dimensional space in terms

Figure 3.7: Complex Morlet Wavelets with different frequencies and bandwidths

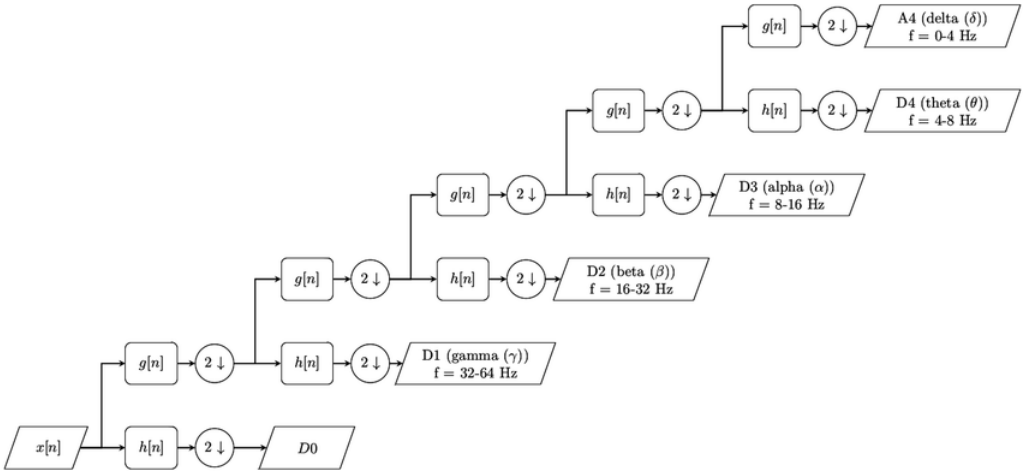


of arousal (the intensity of the emotion) and valence (how pleasurable the emotion is). Instead of using discrete tags, emotions are defined as continuous values between 0.0 and 1.0 where valence represents stronger positive emotions and arousal represents stronger emotional intensity. This approach offers a comprehensive, more nuanced, and psychology informed description of human emotions, thus rendering the model easier to comprehend and applicable to daily life.

In this study, three machine learning and deep learning models, namely Support Vector Regression (SVR), Long Short-Term Memory (LSTM), and a hybrid Convolutional Nueral Network + Long Short-Term Memory (CNN+LSTM) were used to classify emotional states based on features extracted from EEG signals. The models were implemented in Python using libraries such as NumPy, Pandas, Scikit learn, TensorFlow, and Keras.

- **SVR:** Is the regression counterpart of Support Vector Machine. Unlike traditional regression that minimizes squared error, SVR focuses on keeping the predicted values within an epsilon tube around the true values.

Figure 3.8: A 5level DWTbased decomposition of an electroencephalographic (EEG) signal



- **LSTM:** A type of recurrent neural network (RNN) capable of learning long term dependencies in time series EEG data.
- **CNN+LSTM:** A hybrid deep learning architecture where CNN layers were used to extract local spatial features from EEG input, which were then passed to LSTM layers for sequential learning and classification.

Three complementary metrics were computed on the held out test set to assess the regression performance of the arousal and valence models: Mean Absolute Error (MAE), Coefficient of Determination (R^2), and the Pearson Correlation Coefficient (**r**). Together, they quantify prediction error magnitude, explained variance, and linear association, respectively.

Mean Absolute Error (MAE)
MAE measures the average magnitude of errors in a set of predictions, without considering their direction. It is defined as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

where y_i is the true value, $\hat{y}_i$ is the predicted value, and N is the number of samples. A lower MSE indicates that predictions are closer to the actual arousal/valence levels. Unlike MAE, in MSE larger errors are penalized much more heavily which would be good for EEG data as there are so many outliers.

Mean Squared Error (MSE)
Mean squared error (MSE) is a metric to calculate the average of the square of the differences between predicted and actual values of the data. It is determined by taking the average of the squared residuals, where residual = predicted value - actual value for each data point. It can be given by:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$$

where y_i is the true value, $\hat{y}_i$ is the predicted value, and N is the number of samples. A lower MAE indicates that predictions are closer to the actual arousal/valence levels.

Pearson Correlation Coefficient (r)
Pearsons **r** measures the linear correlation between predicted and true values:

$$\mathbf{r} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where x_i is the model's predicted value for the i-th sample, y_i is the true target value (ground truth) for the same sample, and $\bar{x}$ *and* $\bar{y}$ are the mean values of predictions and true labels, respectively. In Pearson Correlation Coefficient values range from -1 to +1, where $\mathbf{r} = +1$ indicates perfect positive linear correlation, $\mathbf{r} = 0$ indicates no linear correlation, and $\mathbf{r}$ = -1 indicates perfect negative linear correlation.

These metrics were computed separately for arousal and for valence, yielding a clear picture of each models predictive accuracy (via MAE), explanatory power (via R^2), and linear agreement with the ground truth (via Pearson **r**).

Support Vector Regression(SVR)
SVR is a regression technique derived from Support Vector Machines (SVM). While SVM is widely used for classification tasks, SVR is designed to predict continuous numerical values. The fundamental objective of SVR is to find a function that best approximates the relationship between the input features and output values, while maintaining an acceptable level of prediction error defined by a margin of tolerance, known as epsilon (ϵ).

Unlike traditional regression methods that attempt to minimize the error for all data points, SVR introduces an epsilon insensitive zone around the predicted function, often visualized as a tube around the regression line. Predictions that fall within this tube are considered sufficiently accurate, and no penalty is applied to them. Only data points that fall outside this epsilon margin contribute to the error term and influence the learning process. These critical points are referred to as support vectors, and they define the final models shape and position.

Mathematically, the model aims to fit a linear function of the form

$$y => f(x) = wx \ + \ b \quad \text{(equation of hyperplane)}$$

Where w is the weight vector, x is the input vector, and b is the bias term. Epsilon insensitive zone or the loss function can be shown as below:

$$-\epsilon \ \leq \ y_i - (wx_i + b) \ \leq \ \epsilon$$

The optimization seeks to keep this function as flat as possible, thereby minimizing the models complexity, while also penalizing large deviations from the epsilon margin using slack variables. These slack variables allow the model to tolerate some data points falling outside the margin if necessary. A regularization parameter $\mathbb{C}$ is used to balance the trade off between model flatness and tolerance to error beyond epsilon.

The optimization goal of SVR becomes minimizing the following objective function:

$$\frac{1}{2}||w||^2 \ + \ \mathbb{C}\sum(\epsilon_i + \epsilon_i^*)$$

Where $||w^2||$ ensures flatness(model simplicity), $\mathbf{C}$ is the regularization parameter and $\epsilon_i,\ \epsilon_i^*$ are slack variables for errors beyond ϵ.

In terms of decision boundaries, SVR creates two lines parallel to the regression line, distanced by epsilon above and below it. The aim is to fit the regression line such that the majority of the data points lie within this band. Any points outside this zone are considered as violations and are penalized proportionally during training.

One of the notable strengths of SVR is its ability to produce robust predictions, especially in cases where the data contains noise or outliers. By ignoring small fluctuations within the epsilon tube and focusing only on significant deviations, SVR can generalize well and avoid overfitting. This makes it particularly useful in real world regression tasks where perfect precision is not required, but stable and reliable predictions are crucial.

The performance of the Support Vector Regression (SVR) model was evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Pearson correlation. The SVR model achieved an MAE of 0.1862 for arousal and 0.1931 for valence. The corresponding MSE scores were 0.0521 (arousal) and 0.0565 (valence), indicating a relatively low average prediction error. The Pearson correlation coefficients were 0.0973 for arousal and 0.2038 for valence, suggesting a weak but noticeable linear relationship between the predicted and actual values.

Long Short Term Memory(LSTM)
LSTM networks are a special type of Recurrent Neural Network (RNN) designed to handle sequential data and learn long term dependencies. Traditional RNNs struggle with the vanishing gradient problem when dealing with long sequences, making it difficult to retain information from earlier time steps. LSTMs overcome this limitation through a unique memory cell structure that includes three gates: the forget gate, which decides what past information to discard; the input gate, which determines what new information to store; and the output gate, which decides what information to output based on the cell state. These gates allow the model to maintain, update, and pass information over long sequences effectively.

Forget Gate
The forget gate decides what information from the previous cell state should be discarded:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

Input Gate
The input gate determines what new information will be stored in the cell state:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

Cell State Update
The cell state is updated using the output of the forget and input gates:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Output Gate
The output gate determines what the next hidden state should be:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh(C_t)$$

Where,

- σ is the sigmoid activation function.
- $*$ denotes element wise multiplication.
- x_t is the input at time step t.
- h_t is the hidden state at time t.
- C_t is the cell state at time t.
- W and b are the weights and biases for each respective gate.

In the context of EEG based emotion recognition, LSTMs are ideal because EEG signals are inherently temporal and they change over time and often contain subtle patterns that span multiple seconds. An LSTM model can capture the progression of brain signals and relate them to emotional states such as arousal and valence.

The LSTM model that was also implemented and evaluated using the same metrics yielded an MAE of 0.1649 for arousal and 0.1993 for valence. The MSE scores were 0.0425 and 0.0577 respectively. The Pearson correlation coefficients were 0.1226 for arousal and 0.0984 for valence, reflecting the models ability to learn temporal dependencies in the EEG data, though the overall correlation remained modest.

CNN+LSTM
CNN + LSTM models combine the strengths of Convolutional Neural Networks (CNNs) and LSTM networks to effectively learn both spatial patterns and temporal dependencies from sequential data like EEG signals. CNNs are excellent at detecting local patterns in input data such as wave like structures, frequency bursts, or signal peaksby applying learnable filters across the input space. This makes them ideal for feature extraction from raw or engineered EEG signals. On the other hand, LSTMs are well suited to learning from time dependent sequences, helping capture temporal dependencies. When used together, CNNs first reduce the dimensionality and extract meaningful spatial features, and then LSTMs model the sequence of those extracted features to understand the dynamics and temporal relationships in the EEG signals.

The basic operations in a CNN involve convolution, where a kernel K slides across the input signal X, computing a feature map:

$$Conv(X) = X * K + b$$

This is typically followed by a non linear activation function. In this study, ReLU was used:

$$ReLU(x) = max(0, x)$$

and used maxpooling to downsample the signal:

$$MaxPool(x) = max(x_1, x_2, x_3, ...x_n)$$

Once spatial features are extracted, they are passed to the LSTM, which uses the same gates as described earlier (forget gate, input gate, output gate) to learn the sequence of those features across time.

The hybrid CNN+LSTM model demonstrated the strongest correlation among the models tested. It achieved an MAE of 0.2591 for arousal and 0.2374 for valence. Despite slightly higher error metrics (MSE: 0.0950 for arousal, 0.0796 for valence), the Pearson correlation coefficients were 0.2950 for arousal and 0.5109 for valence. These values indicate a more substantial linear relationship between the predicted and true values, highlighting the models effectiveness in capturing both spatial and temporal EEG features.

Among the models evaluated, the CNN+LSTM model demonstrated the strongest correlation with the ground truth labels, particularly for valence, indicating its effectiveness in capturing both spatial and temporal dynamics of EEG data. Interestingly, the SVR and LSTM models exhibited lower prediction errors, which may be attributed to the relatively small sample size, simpler models like SVR or a shallower LSTM may generalize better in such scenarios. Nonetheless, the CNN+LSTM's higher correlation suggests it holds promise for generating more emotionally accurate predictions when trained on larger datasets.

3.2 Music Generation

3.2.1 Data Collection

Initially, the goal was to perform transfer learning on a pre trained melody generation model, conditioned on arousal and valence values. To support this, a dedicated study was conducted using the same set of 104 songs (39 English, 65 Sinhala) from the EEG experiment. In this study, 39 participants each annotated 8 songs, including their full versions (melody + vocals), isolated vocals, and isolated melodies. Each clip was annotated by three different participants to minimize bias in emotional labeling.

The intention was to use the 104 melody annotations to pre train a model capable of emotion conditioned melody generation. However, preliminary outcomes revealed several limitations, including the relatively small dataset size for effective transfer learning, lack of access to suitable pre trained models compatible with the annotation format, and computational constraints. As a result, the transfer learning approach could not be pursued.

Given the broad scope and limited time available, an alternative solution was adopted: mapping arousal valence values directly into textual emotion descriptions (e.g., "happy

and energetic", "calm and sad"). These descriptions were then used as prompts for the MUSICGEN ((Copet et al. 2023)) model, a state of the art text to music generation system. As supported by the literature review, MUSICGEN is capable of generating musically coherent outputs from high level emotional text prompts, making it a suitable alternative for the task of emotion reflective melody generation.

3.2.2 Text to Melody Generation

Arousal and valence values from the labeled dataset were initially brought into alignment and examined to observe how emotional states are distributed among the songs. These continuous values were then converted into some emotional categories with the help of a quadrant based model(mentioned in the section 4), which is commonly utilized in the research of emotions. Each quadrant represented a general emotional state as below:

- High Arousal, High Valence "Happy and Energetic"
- High Arousal, Low Valence "Tense and Anxious"
- Low Arousal, High Valence "Calm and Peaceful"
- Low Arousal, Low Valence "Sad and Depressed"

Based on these mappings, each song or melody annotation was assigned a descriptive emotion label in natural language.

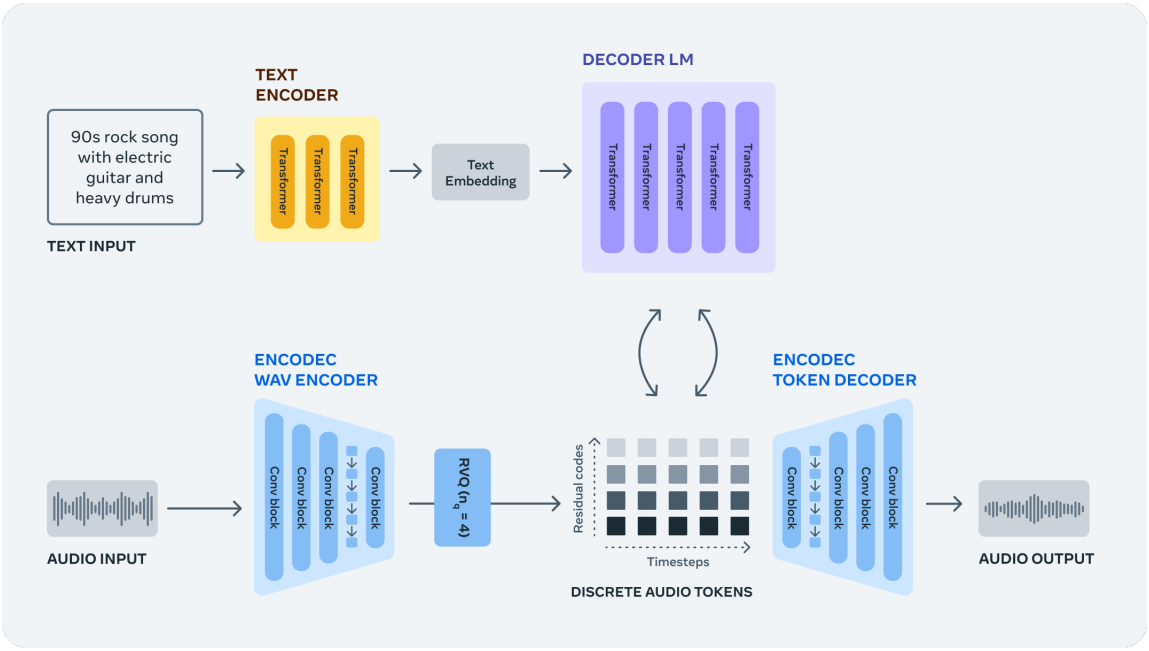
The emotional lables that produced were used as text inputs for the MUSICGEN model. MUSICGEN is a transformer model that can produce expressive and understandable musical pieces from natural language. Using MUSICGEN's ability to understand emotional language, melodies were produced that matched the emotional tone conveyed by the original arousal valence values. Since the model produces music probabilistically, MUSICGEN can produce different musical outputs even with repeated use of the same text input. This is due to random sampling at each step such that there can be numerous different melody variations conveying the same emotional content but varying in musical form, instrumentation, or phrasing. This is particularly helpful in creative work since it introduces diversity and freedom into music composition without the necessity for additional data or new labels.

3.3 Discussion

The processes explored in this study are consistent with prior findings in the domain of EEG based emotion recognition and emotion conditioned music generation (Chapter 2). However, several methodological and practical limitations were identified that need to be addressed to improve the efficiency, accuracy, and applicability of this research area.

One of the key challenges observed was in the subjective annotation of emotional responses to the vocals. While the use of YouTube comments and sentiment analysis helped categorize songs into arousal valence quadrants, the emotional interpretations varied significantly across participants. Despite the effort to average out annotations, individual biases were still present. For instance, the same song was perceived as sad by one participant and peaceful by another. This indicates a limitation in the current

Figure 3.9: MusicGen Model Architecture



annotation approach and suggests the need for either more refined emotional taxonomies or the inclusion of physiological or behavioral correlates to support subjective reporting.

The EEG data acquisition and pre processing pipeline, although thorough and compliant with best practices, was found to be highly resource and time intensive. Each participant session spanned nearly an hour, and subsequent pre processing using EEGLAB including ICA, artifact removal, and re referencing, took several hours per recording. This inefficiency makes it difficult to scale the system for larger studies or real time applications. Similar concerns have been echoed in the literature, underscoring the need for automated or semi automated pipelines that maintain quality while reducing processing time.

In terms of feature extraction, the combination of CWT and DWT provided complementary insights into EEG dynamics. However, CWT was found to be computationally expensive, and DWT, though it is efficient but offered less interpretability in temporal resolution. The dual usage allowed for comparative benchmarking, but future research should focus on optimizing wavelet selection and dimensionality reduction techniques to balance computational efficiency and performance.

For classification, the deep learning models (LSTM and CNN+LSTM) outperformed the traditional SVM approach, especially the Hybrid model, particularly in capturing temporal patterns and spatial features of the EEG signals. However, the small dataset size and limited number of annotations per song may have constrained the models' generalizability. Additionally, a full scale comparison between different deep learning architectures was beyond the scope of this study, indicating the need for more exhaustive algorithmic evaluations in future work.

Regarding the music generation component, the initial plan to use transfer learning with a pre trained melody generation model was limited by dataset size and hardware constraints. The alternative method using arousal valence based text prompts with the MUSICGEN model was more practical and yielded emotionally aligned melodies. However,

since MUSICGENs outputs are probabilistic, the consistency of emotional expression varied between generated samples. This stochastic behavior, while useful for creative diversity, raises questions about how to evaluate the emotional accuracy of AI generated music systematically.

Chapter 4

Implementation and Evaluation

In this section, we'll be explaining the implementation steps done during the EEG Emotion Recognition Framework and Music Generation steps which were discussed in Methodolgy(Chapter 3) and the Evaluation process done during the research study.

4.1 EEG Emotion Recognition Framework

First, The focus will be on the implementation of the Data Acquisition framework, Preprocessing methods, feature extraction and classification models used during the study.

4.1.1 Comment-Based Emotion Estimation of Song Tracks

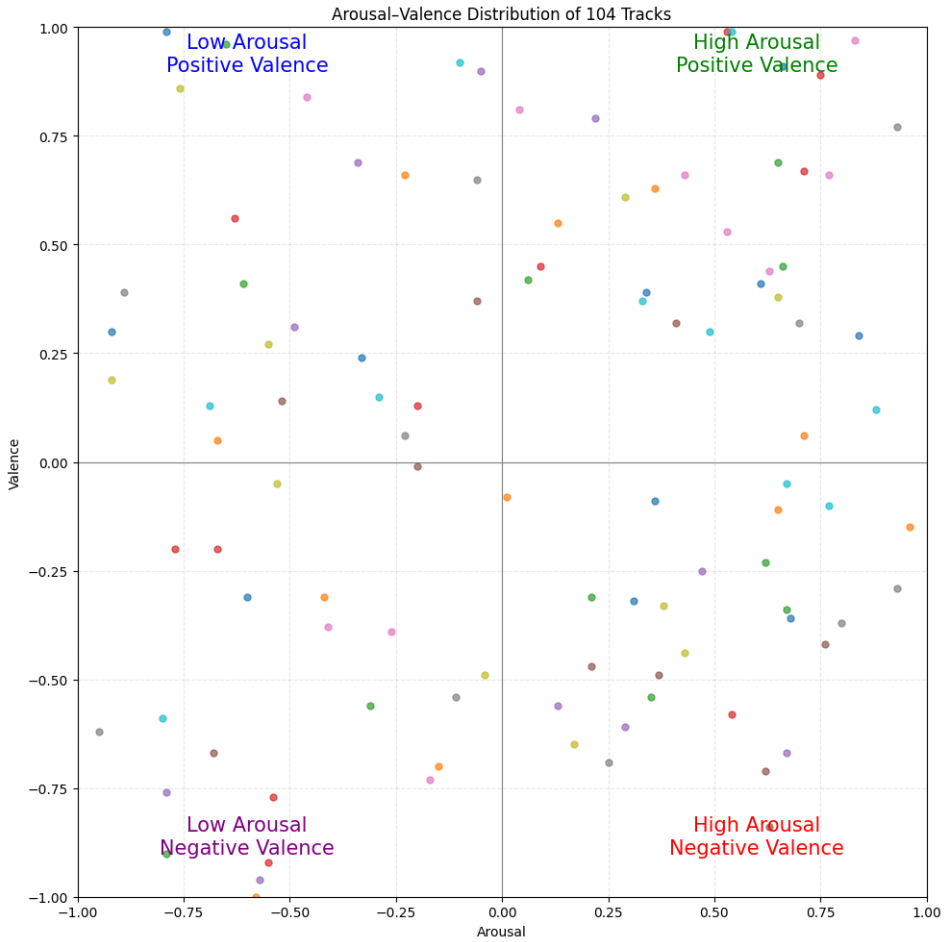
Before conducting the EEG, there was a need to select appropriate songs as the emotional content and diversity of the songs influenced the outcome. To ensure that the selected songs catered for all the emotional responses within the arousal-valence (AV) model, we have developed a specific Python program developed that would analyze user comments on the YouTube video of each song and return some estimation for arousal valence values. People would naturally express how they feel when they listened to a song, so it was a convenient way of getting emotional responses.

The words employed two kindsone for arousal (how much energy or relaxation a song conveys) and one for valence (whether the emotion is positive or negative). For instance, words such as energetic, intense, and fast indicated the higher arousal, whereas words such as calm, relaxing, and chill indicated lower arousal. Similarly, words such as happy, beautiful, and uplifting indicated positive valence, whereas words such as sad, painful, and depressing indicated negative valence.

Each of the comments had their occurrences of certain words verified, and a record of how many times each of them occurred was made. The totals were scaled by the total of comments so songs with lots of comments couldn't skew the results. Each song had then automatically assigned one of the four AV categories. Table 4.1 and figure 4.1 shows the selected song track distribution results and distribution on A-V scale respectively.

This created a quick and easy method for labeling the emotions within the songs prior to their being tested on listeners. It ensured the song set varied in emotions and stayed balanced, as required for training as well as testing emotion recognition systems. Using

Figure 4.1: Track Distribution on A-V Scale



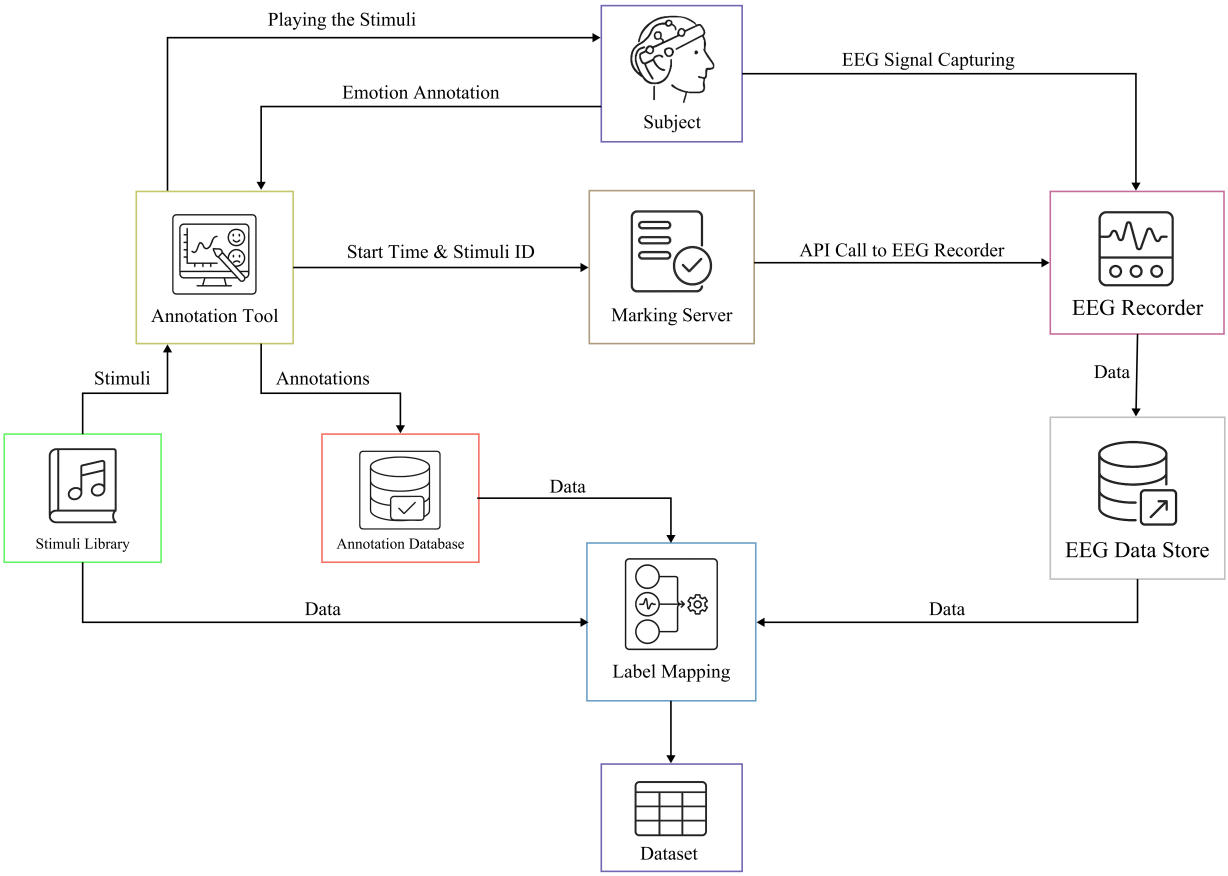
Dimension	Track Count	Percentage
High Arousal - Positive Valence	30	28.8%
High Arousal - Negative Valence	28	26.9%
Low Arousal - Positive Valence	24	23.1%
Low Arousal - Negative Valence	22	21.2%

Table 4.1: Stimuli Selection - Track Distribution Results

listeners' reaction that anyone could view, it provided additional credibility for the feedback from participants on how they themselves felt, making the emotion labels within the study more accurate.

After selecting the music tracks, the Spleeter model was used to separate the vocals and melody, as only the vocal track was required for the EEG data collection experiment. Spleeter is an open-source source separation library developed by Deezer, written in Python and built on TensorFlow. For this study, the 2-stem model was applied, which splits the audio into two components: vocals and accompaniment (melody). This method ensured that only vocal elements were presented during the experiment, minimizing the impact of instrumental components on EEG recordings.

Figure 4.2: Implemented System Architecture



4.1.2 Data Collection Framework

A web User Inteface for data collection of EEG was created after splitting the vocal tracks from the selected music with the Spleeter model. The web application provided an interactive and user-friendly method of listening to vocals and providing feedback. The web page consisted of various important wep pages: an information-gathering page, an instructions page, a stumili interaction page, arousal annotation pages and valence annotating pages, and an Completion page.

The Information Gathering Interface(figure 4.3) gathered data from participants, such as name, gender and age, and provided a brief introduction to the experiment. The Instruction Page(figure 4.4) defined what was to be annotated and provided clear explanations of the terms "arousal" and "valence" as well as an annotated tutorial for annotating and a consent form. At the Stimuli Interaction Page(figure 4.5) , the system randomly selected an available music track from the database and loaded it into the player where participants could play, pause, and replay tracks at their convenience.

In annotation process it followed Self Assessment Manikin(SAM) method in the Arousal Annotation Interface(figure 4.6) enabled people to indicate the extent to which they were excited on a scale. The Valence Annotation Interface(figure 4.7) enabled people to indicate the postivity-negativity of the feelings they had towards the song played to them. The Completion Page informed them that annotation and the recording was complete,

Figure 4.3: Information Gathering

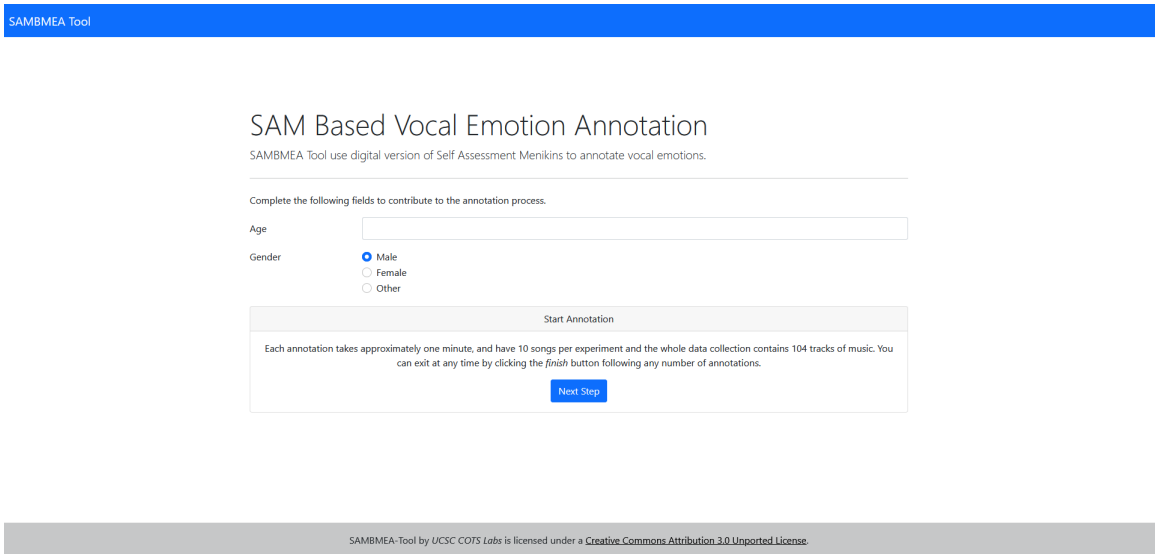
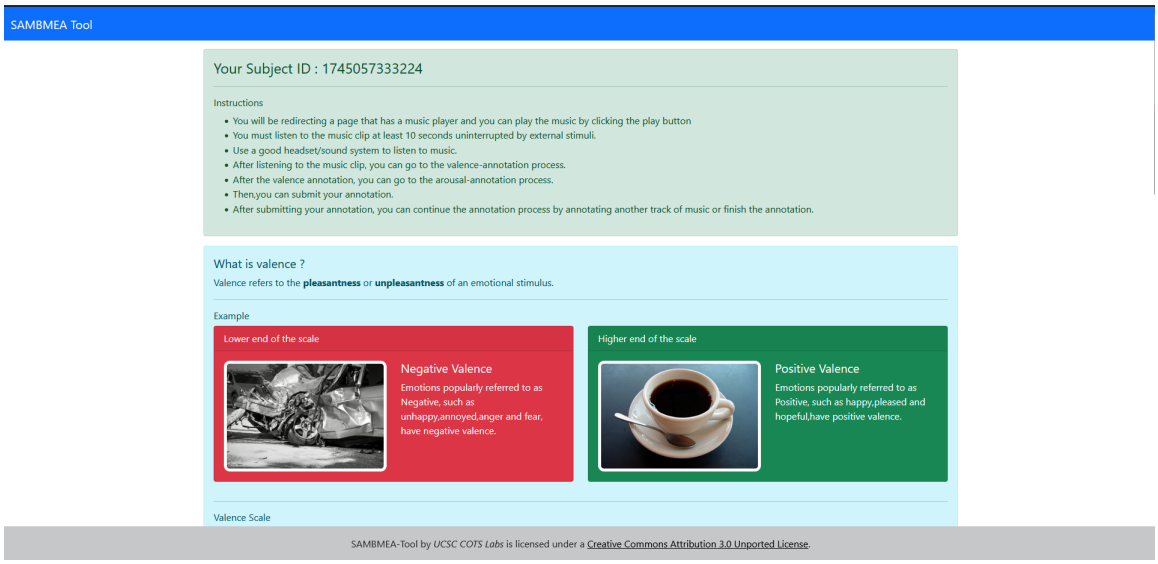


Figure 4.4: Instruction Page



and either they could complete or begin another annotation process.

The front end was developed using JavaScript frameworks, as well as CSS. The back end was created using Node.js and ExpressJS.

EEG recording headset and software was established before the annotation sessions. A 32-channel "EMOTIV EPOC FLEX" system was used to record the EEG signals. The EEG was recorded utilizing Emotiv Pro software that possessed numerous tools for connecting the hardware to the computer and data management. To synchronize the music stimuli with the EEG recordings, there was an added trigger marking server(figure 4.8) to the system. The server provided the starting time of the stimulus to the EEG recorder so that both sound playback and EEG signal could be synchronized.

EEG data was recorded at 128 Hz. The EEG system utilized good Ag/AgCl sensors, and EEG gel was applied to the contact points to reduce resistance. The sensors were placed according to the international 10-20 system. The experiments were done within an

Figure 4.5: Stimuli Interaction Page

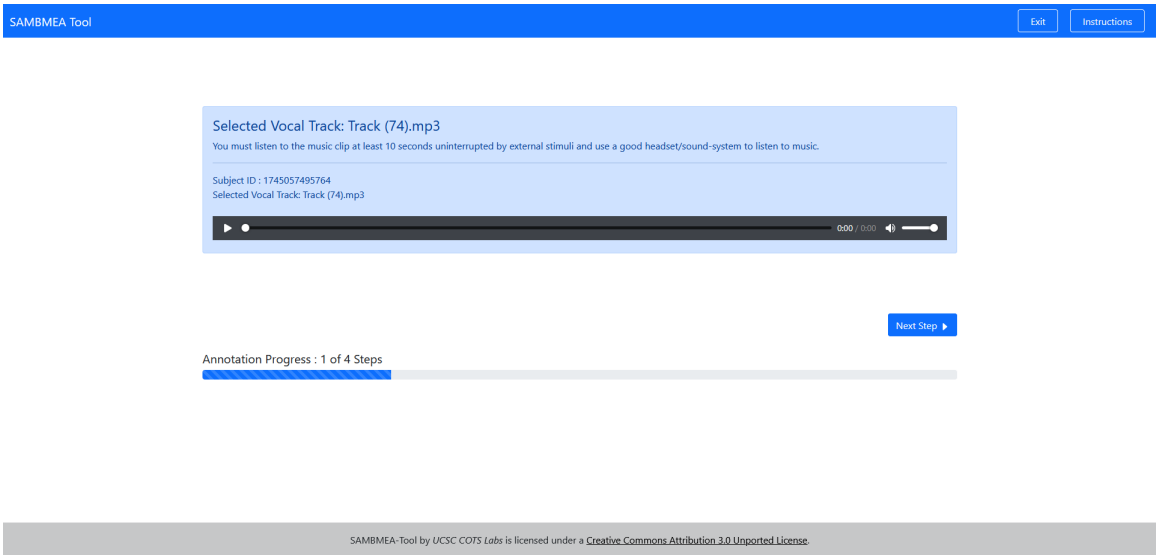
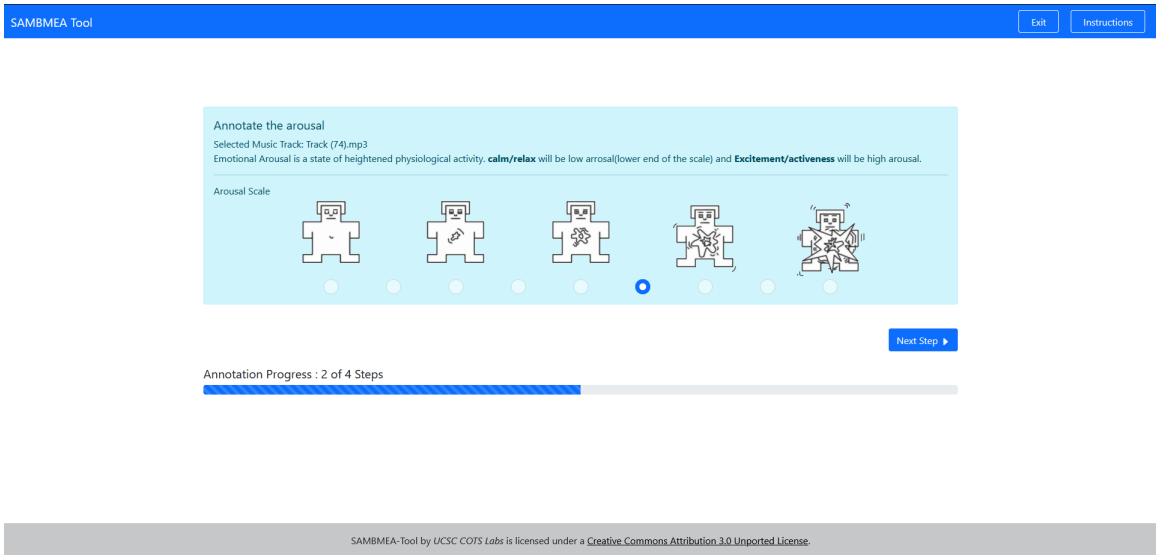


Figure 4.6: Arousal Annotation Interface



isolated laboratory setting(In a Studio), and music was played via Sony SMS-1P near-field studio monitor speakers.

Previously, they recorded the emotions they experienced while undergoing music tests on paper or filled up questionnaires on the computer. They even managed the EEG recording themselves. They initiated and terminated the recording independently and stored every EEG record individually. As a result, it became tedious and time-consuming to prepare the data after the test.

A fully digital annotation system was proposed to address these issues and was implemented(Shown in figure 4.2). The system linked directly the annotation tool to the EEG recorder and possessed multiple key elements: an EEG event extraction and mapping algorithm, an EEG data storage, an annotation database, an EEG recorder, an annotation interface, a music library, and a marking server.

For every session, the annotation tool randomly selected a music track to play indepen-

Figure 4.7: Valence Annotation Interface

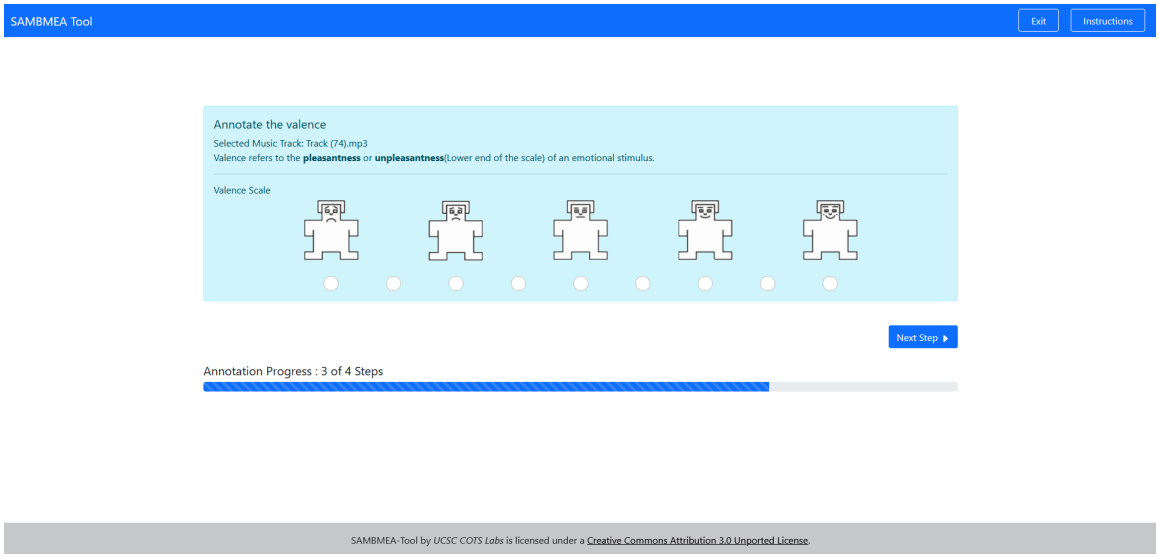
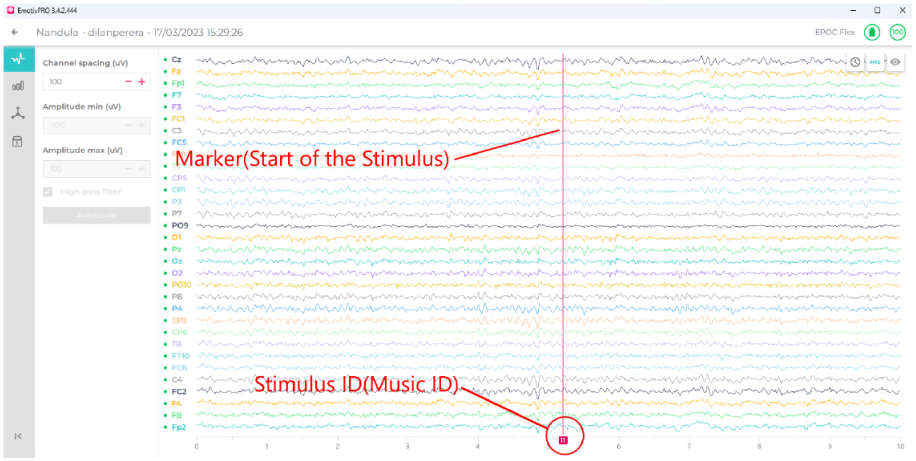


Figure 4.8: Event Marker in EEG Recorder



dently. As the music played, the tool transmitted the stimulus ID and start time to the marking server. The server converted this data into code and transmitted it to the EEG recorder as a signal via Emotivs API and Python serial communication. The participant rated arousal and valence values after listening to the vocal via the interface. The EEG records with markers and the corresponding notes were stored in independent databases.

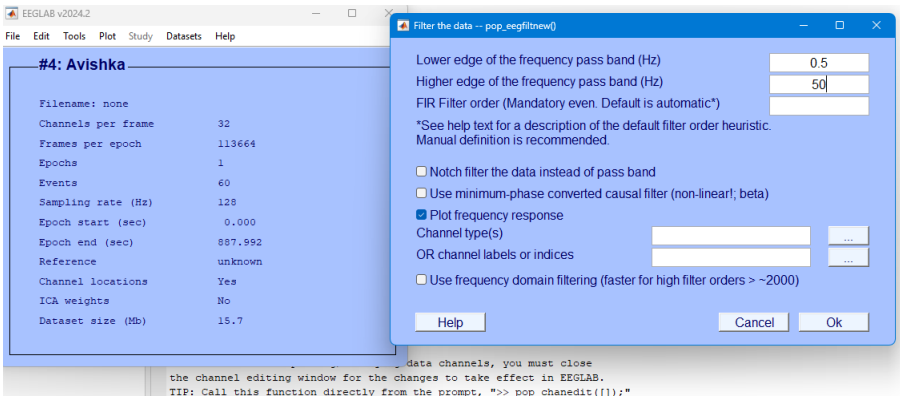
The recorded data passed through an event finder system that labeled events. The system separated EEG signals with markers and linked them to certain notes. The number of additional EEG files was significantly lessened, unwanted data (such as EEG without emotional content) was eliminated, and it became easier to label, reducing the requirement to check manually. The server was created with Flask, a lightweight Python web tool. The server was programmed to receive stimulus events and interact with the EEG hardware so that the given sounds corresponded exactly with the recorded EEG signal. Figure (figure 4.8) is an illustration of the complete system setup, including both the software interface as well as the integration of the markers.

4.1.3 Pre-processing Methodology

The pre-processing step is important for enhancing the quality and reliability of EEG signals, as raw EEG data is often contaminated with various types of noise and artifacts that can compromise the accuracy of downstream analysis. In this study, a comprehensive pre-processing steps was applied using EEGLAB (Delorme and Makeig, 2004), a MATLAB-based toolbox specifically designed for processing EEG data.

The pre-processing steps began by configuring the selected EEG channels and their corresponding scalp locations. Event markers were identified within the data to facilitate accurate segmentation of EEG signals in relation to the onset of stimuli.

Figure 4.9: Finite Impulse Response (FIR) Bandpass Filter Configurations



To eliminate low-frequency drifts and high-frequency noise, a Finite Impulse Response (FIR) bandpass filter was applied, filtering the data between 0.5 Hz and 50 Hz(Shown in figure 4.9). This range was chosen to preserve frequency components relevant to cognitive and emotional processes while removing undesired background activity. In addition, line noise and periodic electrical interference were effectively removed using the CleanLine algorithm(Figure 4.10), configured with the following parameters:

- Remove Channel if it is flat for more than: 50 seconds
- Max acceptable high-frequency noise standard deviation: 10
- Min acceptable channel correlation threshold: 0.5
- Max acceptable 0.5 sec window standard deviation: 25

Following line noise removal, re-referencing was performed to standardize the EEG signals across all channels. For this study, the re-referencing was done by computing the average reference for all channels. This step helps minimize the influence of reference electrode placement and improves comparability across trials and subjects.

Subsequently, Independent Component Analysis (ICA) was applied to decompose the EEG data into statistically independent components. This was done by calling the runica() function with proposed rank of 31(Figure 4.11). The "proposed rank" refers to the number of independent components the runica algorithm determines to be present in the data. Here proposed rank was influenced by the Number of Channels where the algorithm can estimate the rank based on the number of channels or a predefined number of components to compute. These components were automatically classified using

Figure 4.10: CleanLine Algorithm Configurations

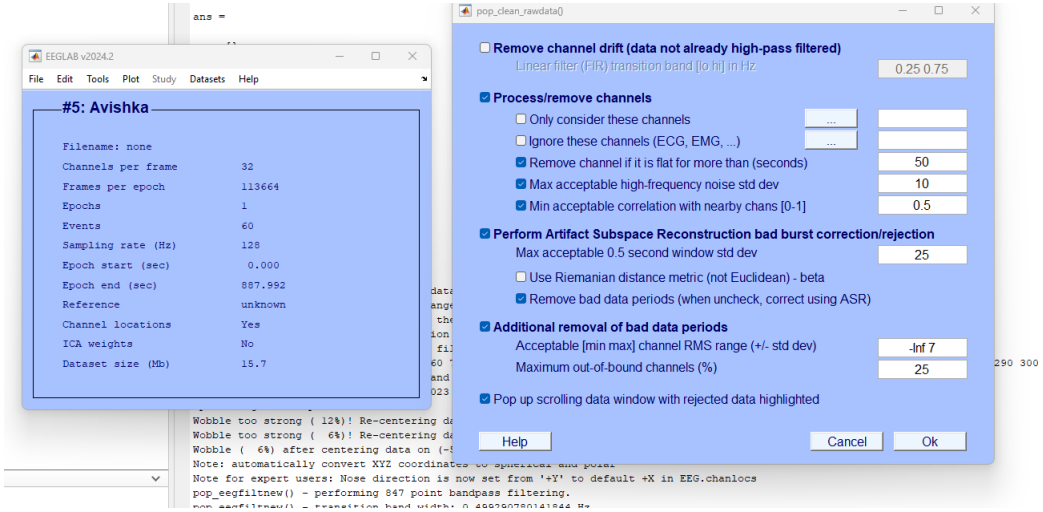
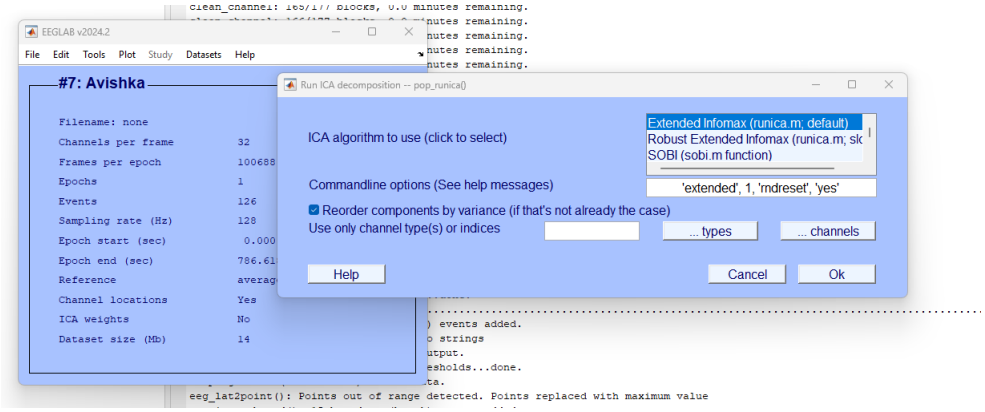


Figure 4.11: Excuting Independent Component Analysis(ICA)



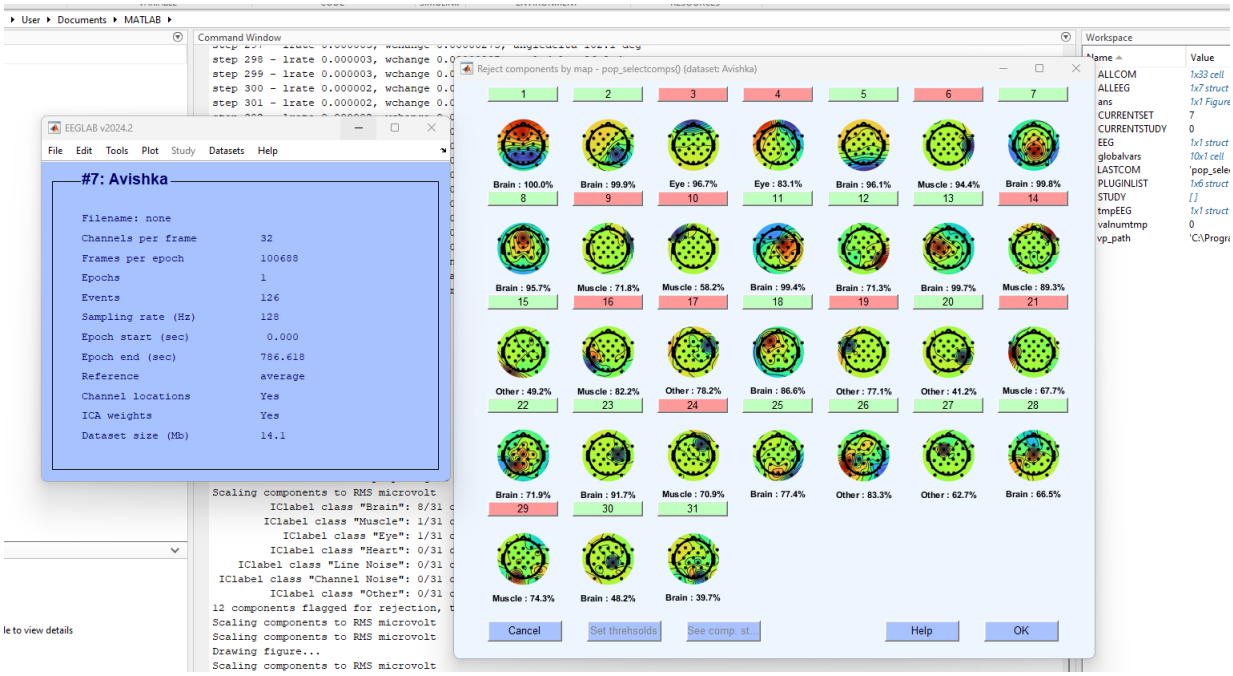
the ICLabel plugin, which assigns probability scores indicating whether a component originates from brain activity, ocular artifacts, muscle noise, line noise, or channel noise. Components identified as non-neural (e.g., eye movements, muscle contractions, or electrical interference) were flagged(Figure 4.12) and removed based on confidence thresholds provided by ICLabel. The parameters for flagging component for rejection was configures as follows:

- Probability range for "Brain": 0.0 - 0.1
- Probability range for "Muscle": 0.9 - 1.0
- Probability range for "Eye": 0.9 - 1.0

Additionally, any boundary events or discontinuities in the EEG recordings were detected and excluded to ensure that only clean and continuous signal segments were retained for further processing(Boundary Events Shown in Figure 4.13).

This multi-stage pre-processing steps substantially improved the signal-to-noise ratio, removed irrelevant and noisy components, and ensured that only high-quality data were retained. The use of EEGLAB, in conjunction with CleanLine and ICLabel, provided a robust and semi-automated approach for artifact removal and signal standardization. Final step is to remove epocs from the dataset for feature extraction. Removing epocs

Figure 4.12: Identified Flagged components of ICLabel



means one data comprises of many tracks(10 tracks in our study), so have to extract cleaned eeg recording for each track before moving on to feature extraction step. Finally, the cleaned EEG dataset obtained was both reliable and consistent, making it well-suited for subsequent stages of feature extraction, emotion annotation alignment, and classification analysis. In the two figures shown below 4.14 and 4.15 you can see the difference between the EEG signals before and after pre-processing.

4.1.4 Feature Extraction Implementations

Feature extraction was performed using both Continuous Wavelet Transform (CWT) and Discrete Wavelet Transform (DWT). The pre-processed EEG data, originally stored in EEGLAB format (.set files), was loaded using an MNE-Python library capable of handling EEG datasets. To ensure accessibility for downstream analysis, the data was fully loaded into memory during the import process. Upon loading, key metadata was extracted, including the sampling frequency, which defines the temporal resolution of the EEG signal, and the total number of recorded samples. These parameters were used to calculate and verify the overall recording duration.

To standardize the input length across all samples for consistent feature engineering and dimensionality reduction, each EEG recording was cropped to a fixed duration of 10 seconds. This ensured uniformity in data length, which is critical for feeding into machine learning models. Following the cropping process, the EEG data was converted into a two-dimensional matrix, where each row corresponds to an EEG channel and each column represents a time point.

The core of the transformation pipeline involved applying a Continuous Wavelet Transform to each EEG channel individually. This transformation was carried out using a Morlet wavelet, a complex wavelet that provides a balanced resolution in both time and

Figure 4.13: Boundary Event Removal

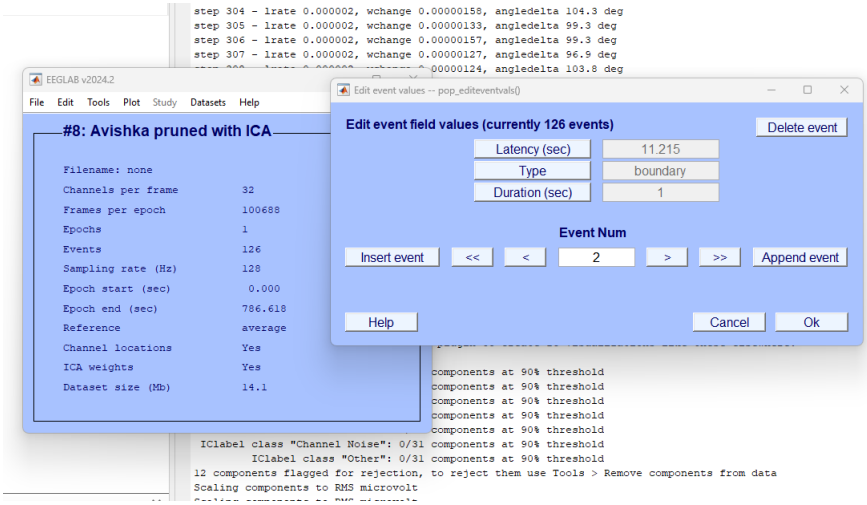
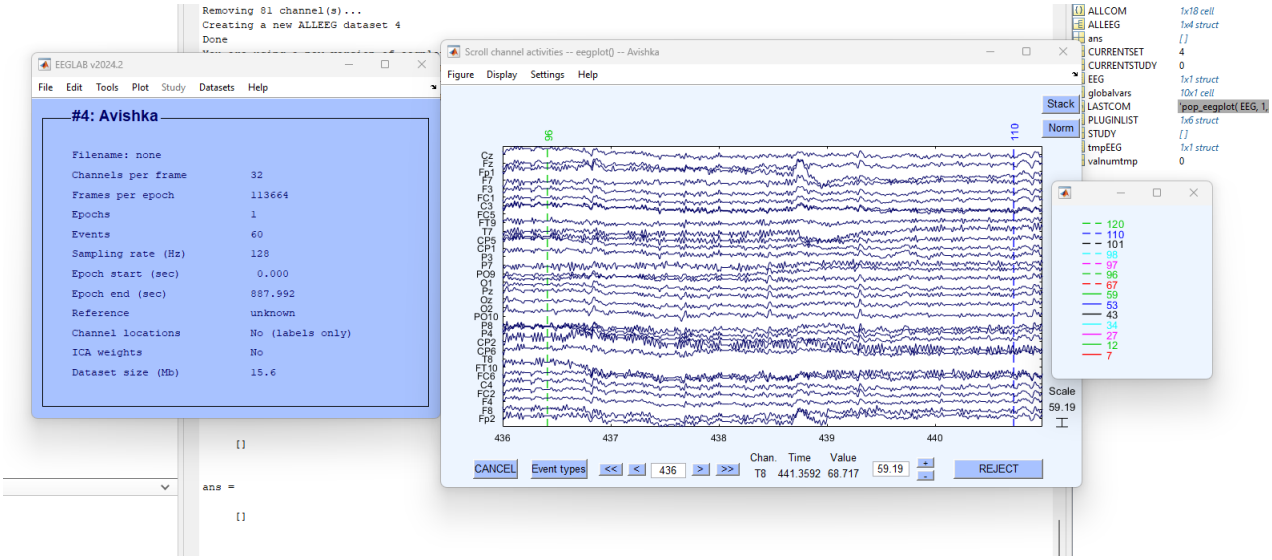


Figure 4.14: Before Pre-Processing

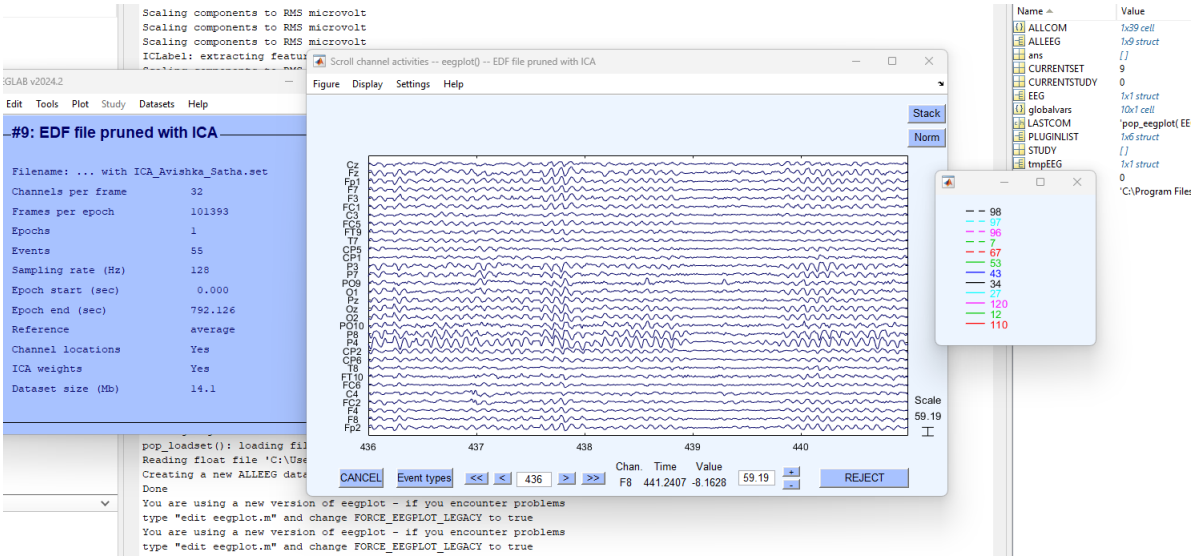


frequency domains, making it well-suited for analyzing non-stationary EEG signals. The signal was decomposed over 64 scales, allowing the analysis of time-varying frequency components across a wide spectrum. For each channel, this process produced a time-frequency matrix representing how the frequency content of the signal evolved over time.

The resulting wavelet coefficients for all EEG channels were organized into a three-dimensional array, structured to preserve information across spatial (channel), frequency (scale), and temporal (time) domains. This representation was particularly well-suited for statistical analysis and deep learning applications, enabling the extraction of both global and localized patterns in the EEG data. It was further utilized as input to neural models such as CNNs and LSTMs, which are capable of capturing complex spatial-temporal dependencies.

For the DWT-based analysis, also like in CWT, 10-second segment was individually processed using Discrete Wavelet Transform to extract hierarchical frequency-domain features. The PyWavelets (pywt) library was used to perform multi-level wavelet de-

Figure 4.15: After Pre-Processing



composition. The db4 Daubechies wavelet was selected as the mother wavelet due to its balance between time and frequency localization and its widespread use in EEG analysis. Decomposition was carried out up to level 5, yielding both approximation and detail coefficients at each level, where each level corresponds to a distinct frequency band. The correspondence between decomposition levels and canonical EEG bands was established based on the sampling rate and the DWTs dyadic scale property.

For each decomposition level, the spectral power was computed by taking the squared magnitude of the wavelet coefficients and summing them within each frequency band of interestdelta (0.5 - 4 Hz), theta (4 - 8 Hz), alpha (8 - 13 Hz), beta (13 - 30 Hz), and gamma (>= 30 Hz). These band-specific power features were extracted for each EEG channel and then concatenated to form a comprehensive feature vector representing the entire EEG segment.

This feature vector served as input for downstream machine learning tasks such as classification or regression. Compared to the CWT-based representation, the DWT method offered a more compact and computationally efficient feature set. However, it trades off some temporal resolution and continuity in frequency representation, which may affect performance depending on the nature of the classification task. Nonetheless, comparing both DWT and CWT approaches provided valuable insight into the trade-offs between discrete and continuous wavelet frameworks for EEG-based emotion recognition.

4.1.5 Feature classification Models

The first model architecture implemented in this study is a deep LSTM network that takes in EEG data in the shape of (1281, 960), where 1281 represents time steps and 960 represents features extracted at each step. It begins with a wide LSTM layer of 256 units to capture rich temporal dynamics, followed by batch normalization and dropout for regularization. A second LSTM layer with 128 units continues the temporal abstraction, again with regularization. The third LSTM layer with 64 units summarizes the entire sequence into a compact representation by setting return sequences = False. After the sequence processing, a dense layer with 64 neurons and ReLU activation provides

non-linear feature transformation, and the final dense layer outputs two continuous valuesarousal and valenceusing a linear activation function. The model is compiled with the Adam optimizer, mean squared error (MSE) loss function, and mean absolute error (MAE) as an additional evaluation metric, which are well-suited for this type of regression task. This layered approach ensures the model learns complex patterns in EEG data across time while maintaining robustness and generalizability.

Feature Extraction Method	Emotion Dimension	MAE	MSE	Pearson Correlation
CWT	Arousal	0.1649	0.0425	-0.1226
	Valence	0.1993	0.0577	0.0984
DWT	Arousal	0.1639	0.0423	0.2275
	Valence	0.1977	0.0577	0.0850

Table 4.2: Evaluation Results - Deep LSTM

In the CNN + LSTM model, the input shape corresponds to EEG time series where each timestep includes a set of features (e.g., EEG channels or wavelet coefficients). You begin with a 1D convolution layer with 64 filters and a kernel size of 3, followed by batch normalization to stabilize learning and max pooling to reduce the spatial dimension. Dropout is used for regularization. A second Conv1D layer with 128 filters deepens the spatial abstraction before another round of normalization, pooling, and dropout.

Feature Extraction Method	Emotion Dimension	MAE	MSE	Pearson Correlation
CWT	Arousal	0.2591	0.0950	0.2950
	Valence	0.2374	0.0796	0.5109
DWT	Arousal	0.1613	0.0420	-0.0865
	Valence	0.1980	0.0577	-0.0089

Table 4.3: Evaluation Results - Hybrid Model (CNN+LSTM)

After the CNN layers, we have introduce a single LSTM layer with 64 units. This layer learns to model how the spatial features evolve over time, capturing emotional state transitions. Dropout is again applied to prevent overfitting. Finally, a dense layer with 2 outputs and linear activation predicts the arousal and valence values, making the model suitable for continuous regression tasks. The model is compiled using the Adam optimizer with Mean Squared Error (MSE) as the loss function and Mean Absolute Error (MAE) as a performance metric.

Support Vector Regression (SVR) was chosen considering the small dataset available, consisting only of 169 EEG samples. The method is especially well-suited here, presenting satisfactory performance even with a smaller number of training samples, and with a lower chance of overfitting compared to more complex deep models. The features for predictions were extracted from EEG signals using both Discrete Wavelet Transform (DWT) as well as the Continuous Wavelet Transform (CWT), thus capturing intricate time-frequency features of the emotional states.

The features extracted initially using multidimensional configurations, including time windows, channels, and frequency components, were then converted to one-dimensional vectors to conform to the SVR's input requirements. The emotional labels, namely arousal and valence, were formatted as a two-column matrix where each row described

the emotional state for one sample. Standardization of features was performed before the training step using z-score normalization to guarantee equal scaling on all inputs, a requirement for the proper operation of kernel models like SVR. The data were split between training and test sets at a ratio of 80-20 for evaluating the generalization ability of the model on test data that had not been seen during training.

Also two different Support Vector Regression (SVR) models were trained, one for predicting arousal and one for estimating valence. Each was set to use a radial basis function (RBF) kernel, which is especially suited to finding non-linear relationships between the input features and their respective target values. Hyperparameter selection, including the penalty factor (C) and epsilon within the loss function, followed widely established defaults, but there is a potential for tuning that might lead to improved results. In order to determine how accurate the predictions were, a variety of regression performance measures were implemented. The Mean Absolute Error (MAE) and Mean Squared Error (MSE) both provided insights regarding the magnitude of the prediction errors, while Pearson correlation coefficients were obtained to examine both the magnitude and the direction of the linear relationship between observed values and predicted values.

Feature Extraction Method	Emotion Dimension	MAE	MSE	Pearson Correlation
CWT	Arousal	0.1862	0.0521	0.0973
	Valence	0.1931	0.0565	0.2038
DWT	Arousal	0.2005	0.0610	0.1446
	Valence	0.1949	0.0544	0.2924

Table 4.4: Evaluation Results - SVR

4.2 Music Generation Model

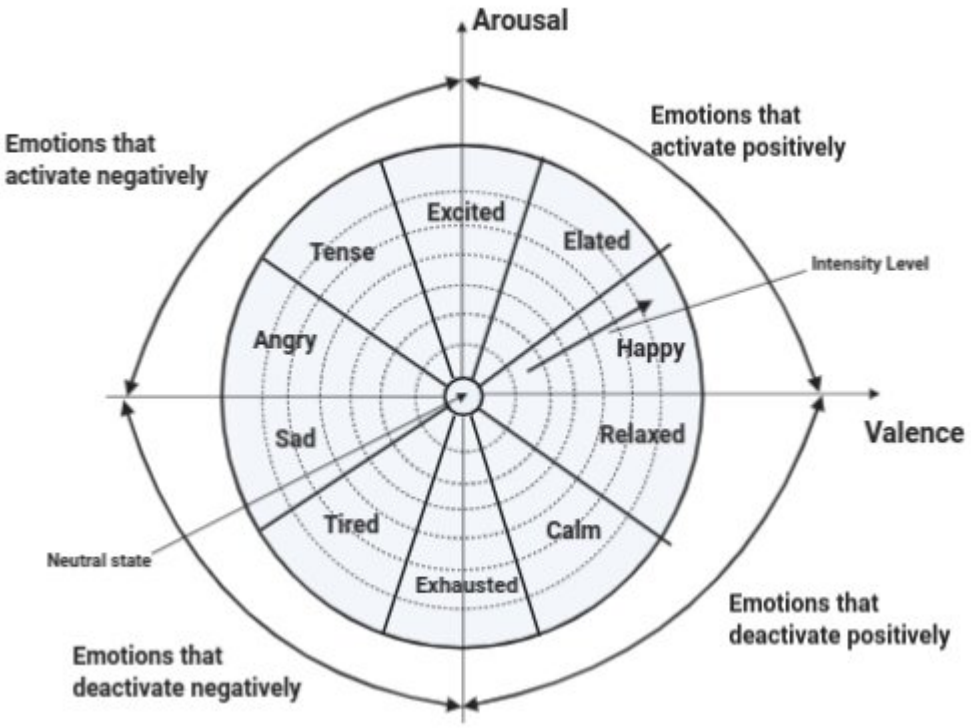
To translate emotions into music, an arousalvalence pair was converted into natural language words by employing a rule-based system. The arousal and valence values were derived from a sample's emotion classifier that was attached to every music sample and were displayed on a graph to observe the way the emotions are distributed across the samples. To make things easier, the continuous emotion scores were divided into emotion labels utilizing an ten-quadrant model. In this quadrant model, the arousalvalence plane was partitioned into ten regions of emotion proposed as in (Sehgal, Sharma & Anand 2021) with varying combinations of emotional arousal and valence. Figure 4.16 shows the Russels circumplex model of emotions mapped to certain label quadrants.

To operate this model, a set of conditional sentences was defined to according to the figure shown in 4.16 to assign natural language prompts to each emotional zone. These prompts were crafted to reflect the musical character and emotional essence of each region.

These descriptive prompts were then used as input to MUSICGEN, a transformer-based language-to-music generation model. The model interprets the emotional cues embedded in the prompt and synthesizes corresponding melodies. Due to MUSICGENs probabilistic generation mechanism, repeated use of the same prompt may yield different musical outputs, enabling diverse emotional expressions under the same emotional label.

To enhance personalization and flexibility, additional user-defined parameters were supported in the prompt formulation. These included melody duration, instrumentation,

Figure 4.16: Regions of Emotion Labels



genre, and tempo, allowing more fine-grained control over the musical output. For instance, users could request: Create a deeply emotional, slow, and sad composition like a lament, using strings and piano, 30 seconds long, in a classical style.

This rule-based mapping approach, combined with prompt augmentation, established a transparent and controllable interface between emotion recognition and generative music synthesis, enabling both systematic evaluation and artistic exploration.

Table 4.5 shows the evaluation results for 10 melodies which was generated by various arousal and valence inputs.To test the quality and emotional fit of the music samples generated, a subjective evaluations was conducted. Each melody was generated from a given arousalvalence pair using a set of rules. The subjects were exposed to the samples and was asked to annotated arousal and valence values and, Music Quality (quality of sound overall, sound, musicality). Music Quality was rated on a continuum from 1 to 10.

In addition to the scores, the listeners were questioned on their opinions regarding the music. What they said illustrated how individuals responded to the music even more clearly than the scores themselves did. It was intended to blend figures with actual emotions to determine how well the system translated emotions into music that individuals could comprehend clearly and feel it themselves(Feedbacks also shown in Table 4.5).

Track No	Avg. Arousal	Avg. Valence	Input Arousal	Input Valence	Avg. Audio Quality	Comments
1	0.20	0.13	0.1	0.2	9.0	Mystical and calm. Good audio quality. Feels sad.
2	0.83	0.33	0.9	0.2	8.0	The melody is intense, but the audio is repetitive. Nothing much to be happy about. Good audio quality.
3	0.60	0.53	0.6	0.6	7.33	Monotonous. Lacks variety. Feels neutral.
4	0.70	0.67	0.9	0.8	7.67	The audio is intense. Good bass. A bit noisy.
5	0.67	0.53	0.5	0.5	7.0	Neutral audio. Sounds like background music suitable for work.
6	0.37	0.53	0.2	0.7	8.33	Soothing and calm music. Less intensive.
7	0.63	0.43	0.8	0.3	8.33	Dominated by the bass. Other instruments are hard to hear.
8	0.60	0.27	0.7	0.1	8.33	Good beat. Feels energetic.
9	0.50	0.70	0.4	0.6	8.33	Good audio quality. Bit of noise present. Gives good vibes.
10	0.40	0.63	0.3	0.7	8.0	Dominated by a humming sound. Repetitive, but gives pleasant vibes.

Table 4.5: Emotion Ratings, Audio Quality and Comments for the Generated Melodies

Chapter 5

Results and Analysis

This study explored two primary feature extraction methods: Discrete Wavelet Transform (DWT) and Continuous Wavelet Transform (CWT) for EEG-based emotion recognition, targeting arousal and valence dimensions. The extracted features were used to train and evaluate three machine learning models: Support Vector Regression (SVR), Long Short-Term Memory (LSTM) networks, and a hybrid Convolutional Neural Network with LSTM (CNN+LSTM). The models were assessed using three evaluation metrics, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Pearson correlation coefficient. This section presents an analysis of the obtained results, comparing model performance across feature extraction methods and reflecting on their alignment with findings in existing literature.

5.1 DWT-Based Results Analysis

In the DWT feature extraction approach, features were extracted using a 5-level decomposition with the Daubechies 4 (db4) wavelet. Band-specific power features were computed for delta (0.5 - 4 Hz), theta (4 - 8 Hz), alpha (8 - 13 Hz), beta (13 - 30 Hz), and gamma (>= 30 Hz) bands across 32 EEG channels.

When evaluated using Support Vector Regression (SVR), the DWT features yielded modest results. The MAE for arousal was 0.2005 and valence was 0.1949, while the Pearson correlations were 0.1446 and 0.2924, respectively. Although these results demonstrate some linear trend alignment especially for valence, the performance remained limited, highlighting the difficulty SVR has with capturing non-linear temporal dependencies.

The LSTM model showed slight improvements over SVR. Using DWT features, it achieved MAE values of 0.1639 (arousal) and 0.1977 (valence), with Pearson correlations of 0.2275 and 0.0850, respectively. This indicates better estimation accuracy but a weaker ability to capture temporal emotional dynamics, especially for valence.

The CNN+LSTM hybrid model achieved the lowest MAE for arousal at 0.1613, but interestingly, it had negative Pearson correlation values (-0.0865 for arousal and -0.0089 for valence), suggesting poor trend alignment despite low error magnitudes. This may indicate overfitting or the models difficulty in generalizing from the simpler DWT features as well as dataset sample being low.

Model	Emotion Dimension	MAE	MSE	Pearson Correlation
SVR	Arousal	0.2005	0.0610	0.1446
	Valence	0.1949	0.0544	0.2924
LSTM	Arousal	0.1639	0.0423	0.2275
	Valence	0.1977	0.0577	0.0850
CNN + LSTM	Arousal	0.1613	0.0420	-0.0865
	Valence	0.1980	0.0577	-0.0089

Table 5.1: DWT Feature Extraction Evaluation Results

While previous literature reports classification accuracies of 80% or higher using DWT, it is important to note that those studies typically frame emotion recognition as a classification task, mapping EEG signals to discrete labels (e.g., "happy", "sad"). In contrast, this study uses a regression-based approach, predicting continuous arousal and valence scores, which is inherently more challenging and realistic. Thus, while the performance may seem modest compared to classification benchmarks, this study represents a necessary step toward more precise and nuanced emotion modeling.

5.2 CWT-Based Results Analysis

The Continuous Wavelet Transform (CWT) offers a denser, more expressive time-frequency representation by capturing frequency content at multiple scales across the full signal. In this study, the Morlet wavelet was used to generate CWT coefficients across 64 scales for each EEG channel. Unlike DWT, which simplifies signals into band powers, CWT retains both time and frequency continuity, making it ideal for learning models that can extract meaningful temporal dependencies.

The CWT features proved more challenging for SVR. The MAE values were 0.1862 (arousal) and 0.1931 (valence), with Pearson correlations of 0.0973 and 0.2038, respectively. Compared to DWT, SVR performance degraded slightly, especially for arousal, likely due to the higher dimensionality of CWT features which SVR could not handle efficiently.

LSTM models showed mixed performance with CWT. While MAE for arousal remained similar (0.1649) and valence slightly higher (0.1993), the correlation values were 0.1226 for arousal and 0.0984 for valence, indicating weaker alignment. These results suggest that while LSTM can handle sequence data, the raw CWT coefficients may require more pre-processing or dimensionality reduction to be effective in this architecture.

However, the CNN+LSTM model demonstrated the best trend prediction performance using CWT features. It achieved Pearson correlation coefficients of 0.2950 for arousal and 0.5109 for valence, the highest correlations across all experiments, showing its superior ability to capture emotional progression. Despite higher MAEs of 0.2591 (arousal) and 0.2374 (valence), the CNN+LSTM model with CWT proved the most capable in learning the nuanced, temporal patterns that underlie emotional states.

These findings are particularly newly discovered because, to our knowledge, no prior studies have used Morlet-based CWT coefficients in EEG emotion recognition. Thus, while the error margins may appear higher, this approach introduces a novel feature rep-

Model	Emotion Dimension	MAE	MSE	Pearson Correlation
SVR	Arousal	0.1862	0.0521	0.0973
	Valence	0.1931	0.0565	0.2038
LSTM	Arousal	0.1649	0.0425	-0.1226
	Valence	0.1993	0.0577	0.0984
CNN + LSTM	Arousal	0.2591	0.0950	0.2950
	Valence	0.2374	0.0796	0.5109

Table 5.2: CWT Feature Extraction Evaluation Results

resentation that captures richer dynamics, laying the groundwork for future exploration and optimization.

5.3 Cross-Method Comparison between CWT and DWT

To evaluate the relative performance of the DWT (Discrete Wavelet Transform) feature extraction method compared to CWT (Continuous Wavelet Transform), we calculated the percentage change in three evaluation metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Pearson Correlation. These percentage changes are computed using the formula:To evaluate the relative performance of the DWT (Discrete Wavelet Transform) feature extraction method compared to CWT (Continuous Wavelet Transform), we calculated the percentage change in three evaluation metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Pearson Correlation. These percentage changes are computed using the formula:

- % Change = ((DWT - CWT) / CWT) X 100 for MAE and MSE.
- For Pearson Correlation, since values can be negative or close to zero, we use % Change = ((DWT - CWT) / |CWT|) X 100 to better capture the directional difference relative to the magnitude of the CWT correlation.

Model	**Emotion**	**MAE (%)**	**MSE (%)**	**Pearson Corr (%)**
SVM	Arousal	+7.68	+17.09	+48.63
SVM	Valence	+0.93	-3.72	+43.46
CNN+LSTM	Arousal	**-37.72**	**-55.79**	**-129.32**
CNN+LSTM	Valence	**-16.56**	**-27.47**	**-101.75**
LSTM	Arousal	**-0.61**	**-0.47**	+285.69
LSTM	Valence	**-0.80**	0.00	-13.57

Table 5.3: Percentage change from CWT to DWT for MAE, MSE, and Pearson Correlation

The interpretation of these results gives us insight into how each feature extraction method impacts model performance. A positive percentage in MAE or MSE indicates that DWT produced higher error, meaning worse performance, whereas a negative value suggests lower error and improved prediction accuracy. For Pearson correlation, a positive percentage means DWT improved correlation with ground truth values, while a negative percentage indicates a decline in the model's ability to capture true emotion trends.

From the analysis, it is evident that DWT improved correlation for the SVM model in both arousal and valence dimensions, despite a slight increase in MAE. For the LSTM model, DWT significantly boosted the Pearson correlation for arousal (+285%), indicating better temporal learning with DWT-based features. However, CNN+LSTM showed a drastic drop in correlation when using DWT, even though the MAE and MSE were lower. This suggests that while DWT may reduce raw error, it might not preserve the necessary temporal-frequency features for deeper networks like CNN+LSTM, which seem to benefit more from the richer representations of CWT.

Overall, DWT seems better suited for simpler models like SVM and LSTM, especially in scenarios prioritizing correlation over raw error. In contrast, CWT is more effective for deep hybrid models like CNN+LSTM, where capturing nuanced time-frequency dynamics is critical for strong performance.

5.4 Emotion Mapping Strategy

The first step in this process involved mapping the continuous arousal-valence values into natural language prompts that could guide a generative music model. The arousal-valence space was divided into eight emotional zones, each representing a unique quadrant or emotional tone (e.g., high arousal + low valence = aggressive or chaotic; low arousal + high valence = calm or soothing). This mapping was implemented according to an ten-quadrant model where the arousalvalence plane was partitioned into ten regions of emotion proposed as in (Sehgal et al. 2021) with varying combinations of emotional arousal and valence.

This discrete partitioning of a continuous space enabled a meaningful translation between the numeric outputs of the EEG model and a text-driven generative process. Each description was carefully curated to reflect the emotional qualities of the target quadrant, ensuring the music generated would be semantically aligned with the emotional experience of the subject.

5.5 Music Generation with MusicGen

The evaluation of arousal and valence across the 10 tracks reveals meaningful patterns in how emotional characteristics are perceived and how they align with the input values. Arousal, which measures the energy or intensity of a track, shows a generally consistent relationship between the input arousal values and the average arousal ratings. For instance, tracks like 2, 4, and 7, which had higher input arousal values (0.8 or above), also resulted in relatively high average arousal scores, indicating that energetic or intense music is being accurately captured by the input-emotion mapping.

In contrast, the valence ratings, indicating emotional positivity or pleasantness show a weaker and more varied relationship. Some tracks, such as Track 6, had a high input valence (0.7) and a reasonably close average valence (0.53), suggesting good alignment. However, others, like Track 8, had a very low input valence (0.1) but a moderately higher average (0.27), and Track 2 had a low input valence (0.2) but was perceived at 0.33 on average. These differences suggest that valence is more subjective and possibly influenced by melodic, harmonic, or stylistic elements that are not always captured through basic

audio features.

The emotional interpretations for each track support these numeric trends. Track 1, for example, is characterized by low arousal and moderate valence, described as mystical and calm, which matches its peaceful emotional profile. Track 2, which has high arousal but low valence, is noted for being intense yet emotionally flatconsistent with an anxious or agitated mood. Tracks like 6 and 10, which have low arousal but higher valence, are described as soothing or pleasant, reinforcing the idea that music can be emotionally positive even if not highly energetic.

From an audio quality perspective, most tracks maintain a score above 8, indicating generally good production value. Even tracks with comments noting noise or repetitionsuch as Track 3 ("monotonous") or Track 10 ("repetitive") still achieved fair emotional ratings, suggesting that minor imperfections do not significantly hinder emotional perception. However, tracks where elements like bass dominance (e.g., Track 7) or lack of variety are present may slightly affect clarity in emotional delivery.

In the table 5.4 shows how the emotional alignment between the input emotion values correlates with emotion values we got from subjective evaluation. This helps quantify how closely the perceived (average) emotion matches the intended (input) emotion. The formula used:

$$Alignment(\%) = (1 - |Average - Input|)\text{»}100$$

This gives 100% if there's perfect alignment (no difference), and drops as the difference increases.

Track	Arousal Alignment (%)	Valence Alignment (%)
1	80.00	86.67
2	92.59	66.67
3	100.00	88.89
4	77.78	83.33
5	83.33	96.67
6	45.71	76.19
7	79.17	56.67
8	85.71	16.67
9	80.00	91.67
10	75.00	91.43

Table 5.4: Emotion Alignment between Input and Evaluated Arousal/Valence

The emotion alignment results reveal that arousal alignment is generally strong, with most tracks achieving over 80% alignment and Track 3 even reaching a perfect 100%. This indicates that the perceived energy or intensity of the tracks closely matches the intended input values, suggesting effective modeling of arousal. In contrast, valence alignment is more varied, with some tracks like 5, 9, and 10 showing high accuracy (over 90%), while others such as Tracks 7 and 8 exhibit significant mismatches. Track 8, for instance, was intended to be very low in valence (0.1) but was perceived much higher (0.27), resulting in only 16.67% alignment. This discrepancy highlights the subjective nature of valence perception and suggests that modeling valence accurately may require

more nuanced features or refined approaches. Overall, while arousal detection appears robust, valence alignment remains an area for improvement.

Chapter 6

Conclusion and Future Work

This study presents a comparative exploration of DWT and CWT feature extraction methods for EEG-based emotion recognition, implemented in a regression based framework predicting continuous arousal and valence scores. The comparison results reveal that while DWT continues to offer reliable, low-error feature representations, CWT introduces a promising new direction through the use of Morlet wavelet coefficients, a contribution not previously addressed in the literature. Using these predicted arousal and valence values, we mapped them to user preferences to generate melodies that reflect the emotional impression conveyed by song vocals. This mapping serves as an intermediate step toward more personalized and emotion-sensitive music experiences.

Among the evaluated models, the CNN+LSTM architecture, when paired with CWT features, demonstrated the highest Pearson correlation particularly for valence indicating its strength in capturing overall emotional trends even when prediction errors (MAE/MSE) were slightly higher. This makes it suitable for real-time or dynamic emotional analysis tasks, including adaptive music generation, neuro-feedback systems, and affect-aware applications. Also have to mention two datasets were created during the study. First one is the annotated dataset of arousal valence values for the vocal, melody and the full song and has 104 tracks. Second one is the EEG Data for the all 104 vocal tracks with at least two recordings per song vocal. Both the datasets have been made publicly available to encourage further research in emotion recognition and its applications in music generation.

To improve generalizability, future work should prioritize expanding the dataset beyond the current 169 samples, incorporating more participants and diverse stimuli. Additionally, integrating dimensionality reduction techniques such as PCA or autoencoders may help in refining CWT features for compatibility with both traditional and deep learning models, while also addressing overfitting concerns.

Performance could also be enhanced by leveraging advanced architectures such as attention mechanisms or transformers, which would allow the model to focus on emotionally salient parts of the EEG signal. Furthermore, combining EEG with other modalities like facial expression analysis, voice tone, or physiological measurements could lead to more robust and comprehensive emotion recognition systems.

Comparing the results of predicting emotions as classes versus predicting continuous values from EEG data reveals a noticeable difference in accuracy. When emotions are

predicted as discrete classes (e.g., high arousal vs. low arousal, positive vs. negative valence), the model often performs better due to the clear, bounded nature of the classes. This makes the classification problem less complex and may lead to more accurate predictions, as the model can focus on distinguishing between distinct groups.

On the other hand, predicting continuous values for emotional states (such as exact arousal and valence scores) introduces additional challenges. Continuous regression requires the model to not only capture the general trend but also predict more precise values, which can be harder due to the variability and noise present in EEG signals. Small errors in continuous predictions may have a more significant impact on evaluation metrics like Mean Absolute Error (MAE) or Pearson correlation, making them appear less accurate.

In terms of music generation, a promising future direction involves moving beyond text-based melody mapping to a fully data-driven generation model. By training deep generative architectures (e.g., Transformers, VAE, or MusicVAE) on a large-scale dataset of music annotated with continuous arousal-valence scores such as DEAM or curated MIDI/audio corporawe could develop a model that learns to produce music directly from emotional cues. This would enable the generation of emotionally expressive and contextually coherent melodies across various genres and tempos, based on learned patterns rather than predefined mappings.

Such a generative model could be integrated with the EEG emotion recognition pipeline for real-time, neuroadaptive music generation. As the system continuously interprets the user's emotional state from EEG data, it could dynamically generate and adapt music in response, creating a deeply personalized and interactive experience. This could have significant applications in personalized music therapy, immersive media, and emotional wellbeing. Future evaluations should include user studies to assess perceived emotional alignment, music quality, and responsiveness, supported by physiological metrics to validate real-time emotional effects. Overall, this approach represents a significant step toward merging human affective states with artificial creativity.

References

Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., Adeli, H. & Subha, D. P. (2018), 'Automated eeg-based screening of depression using deep convolutional neural network', *Computer methods and programs in biomedicine* **161**, 103–113.

Adolphs, R. (2002), 'Neural systems for recognizing emotion', *Current opinion in neurobiology* **12**(2), 169–177.

Baur, D., Steinmayr, B. & Butz, A. (2010), 'Songwords: Exploring music collections through lyrics.', pp. 531–536.

Berardinis, J. D., Cangelosi, A. & Coutinho, E. (2020), 'The multiple voices of musical emotions: Source separation for improving music emotion recognition models and their interpretability'.

Berger, H. (1929), 'Über das elektroenkephalogramm des menschen', *Archiv für psychiatrie und nervenkrankheiten* **87**(1), 527–570.

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Lee, S., Neumann, U. & Narayanan, S. (2004), Analysis of emotion recognition using facial expressions, speech and multimodal information, *in* 'Proceedings of the 6th international conference on Multimodal interfaces', pp. 205–211.

Cartocci, G., Modica, E., Rossi, D., Inguscio, B. M., Arico, P., Levy, A. C. M., Mancini, M., Cherubino, P. & Babiloni, F. (2019), 'Research article antismoking campaigns perception and gender differences: A comparison among eeg indices'.

Chu, H., Urtasun, R. & Fidler, S. (2016), 'Song from pi: A musically plausible network for pop music generation'.
URL: *http://arxiv.org/abs/1611.03477*

Clarke, E., DeNora, T. & Vuoskoski, J. (2015), 'Music, empathy and cultural understanding', *Physics of life reviews* **15**.

Coan, J. A. & Allen, J. J. (2004), 'Frontal eeg asymmetry as a moderator and mediator of emotion', *Biological psychology* **67**(1-2), 7–50.

Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y. & Défossez, A. (2023), 'Simple and controllable music generation'.
URL: *http://arxiv.org/abs/2306.05284*

Craig, A. D. (2009), 'How do you feel - now? the anterior insula and human awareness'.

Damasio, A. R. (1998), 'Emotion in the perspective of an integrated nervous system1published on the world wide web on 27 january 1998.1', *Brain Research Reviews*

26(2), 83–86.
URL: *https://www.sciencedirect.com/science/article/pii/S0165017397000647*

Darwin, C. (1872), 'The expression of emotions in animals and man', *London: Murray* **11**(1872), 1872.

Davidson, R. J. (1990), 'Electrophysiological correlates of happiness and sadness', *Science 290.5490* pp. 2188–2191.

Davidson, R. J. (2002), 'Anxiety and affective style: role of prefrontal cortex and amygdala', *Biological psychiatry* **51**(1), 68–80.

Dias, D. S. & Fernando, T. G. (2019), Komposer - automated musical note generation based on lyrics with recurrent neural networks, Institute of Electrical and Electronics Engineers Inc., pp. 76–82.

Ekman, P. & Friesen, W. V. (1978), 'Facial action coding system', *Environmental Psychology & Nonverbal Behavior* .

Fan, X., Yan, Y., Xiaomin, W., Yan, H., Li, Y., Xie, L. & Yin, E. (2020), Emotion recognition measurement based on physiological signals, pp. 81–86.

Gable, P. A. & Harmon-Jones, E. (2010), 'The effect of low versus high approach-motivated positive affect on memory for peripherally versus centrally presented information.', *Emotion* **10**(4), 599.

Godet, A., Fortier, A., Bannier, E., Coquery, N. & Val-Laillet, D. (2022), 'Interactions between emotions and eating behaviors: Main issues, neuroimaging contributions, and innovative preventive or corrective strategies', *Reviews in Endocrine and Metabolic Disorders* **23**, 1–25.

Gunes, H. & Piccardi, M. (2011), 'Emotion recognition from face images using a hybrid neural network', *Neurocomputing 74* **12-13**, 2141–2151.

Guo, Y., Liu, Y., Zhou, T., Xu, L. & Zhang, Q. (2023), 'An automatic music generation and evaluation method based on transfer learning', *PLoS ONE* **18**.

Gomez, Danuser, P. B. & Grimm, S. (2019), 'Cultural differences in emotional responses to music: A cross-cultural study', *Frontiers in Psychology* **2465**.

Herrmann, C. S., Matthias, H. M. & Andreas, K. E. (2005), 'Gamma activity in human eeg reflects the resolution of attentional demands', *Neuroimage* **24**, 846–856.

Hu, Y., Yang, H., Huang, H. & He, L. (2024), Cross-modal features interaction-and-aggregation network with self-consistency training for speech emotion recognition, International Speech Communication Association, pp. 2335–2339.

James, W. (1884), 'What is an emotion?'.
URL: *https://www.jstor.org/stable/2246769*

Ji, S., Luo, J. & Yang, X. (2020), 'A comprehensive survey on deep music generation: Multi-level representations, algorithms, evaluations, and future directions'.
URL: *http://arxiv.org/abs/2011.06801*

Juslin, P. N. & Västfjäll, D. (2008), 'Emotional responses to music: The need to consider underlying mechanisms', *Behavioral and Brain Sciences* **31**(5), 559575.

Keil, A., Bradley, M. M., Hauk, O., Rockstroh, B., Elbert, T. & Lang, P. J. (2002), 'Large-scale neural correlates of affective picture processing', *Psychophysiology* **39**(5), 641–649.

Knyazev, G. G. (2013), 'Eeg correlates of self-referential processing', *Frontiers in human neuroscience* **7**, 264.

Kollias, D., Nicolaou, M. A., Kotsia, I., Zhao, G. & Zafeiriou, S. (2017), Recognition of affect in the wild using deep neural networks, Vol. 2017-July, IEEE Computer Society, pp. 1972–1979.

Kreibig, S. D. (2010), 'Autonomic nervous system activity in emotion: A review'.

Krishnan, G. (2023), 'How musicians translate feeling into sound the process and end product (explained by rachel claudio)'.
URL: *https://medium.com/@gauravkrishnan/how-musicians-translate-feeling-into-sound-the-process-end-product-explained-by-rachel-fc922ed2c554*

Levenson, R. W. (2014), 'Whats so cool about cool? childrens use of cognitive strategies to control cool and hot cognition', *Current Directions in Psychological Science* **23**, 281–286.

Madhok, R., Goel, S. & Garg, S. (2018), Sentimozart: Music generation based on emotions, Vol. 2, SciTePress, pp. 501–506.

Mauss, I. B. & Robinson, M. D. (2009), 'Measures of emotion: A review'.

Moncrieff, M. A. & Lienard, P. (2018), 'Moral judgments of in-group and out-group harm in post-conflict urban and rural croatian communities', *Frontiers in Psychology* **9**.

Nuri, A., Niazi, I. K. & Guger, C. (2019), 'Eeg-based emotion recognition: a review of recent applications and future prospects', *Journal of neural engineering* **16**.

Pell, M. D. & Kotz, S. A. (2011), 'On the time course of vocal emotion recognition', *PLoS ONE* **6**.

Pessoa, L. (2013), *The cognitive-emotional brain: From interactions to integration*, MIT press.

Phelps, E. A. (2006), 'Emotion and cognition: Insights from studies of the human amygdala', *Annual Review of Psychology* **57**, 27–53.

Picard, R. W. (1997), *Affective computing*, MIT press.

Plass-Oude Bos, D. (2012), 'Automated artifact detection in brainstream an evaluation of an online eye and muscle artifact detection method'.

Pooja, Pahuja, S. K. & Veer, K. (2022), 'Recent approaches on classification and feature extraction of eeg signal: A review'.

Roberts, A., Engel, J., Raffel, C., Hawthorne, C. & Eck, D. (2018), 'A hierarchical latent vector model for learning long-term structure in music'.
URL: *http://arxiv.org/abs/1803.05428*

Russell, J. A. & Barrett, L. F. (1999), 'Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant', *Journal of Personality and Social Psychology* **76**.

Schmidt, L. A. & Trainor, B. C. (2001), 'Frontal eeg asymmetry and the behavioral activation and inhibition systems.', *Psychophysiology* **38**, 847–855.

Schuller, B. (2013), 'Speech emotion recognition: Features, databases, and challenges', *IEEE Transactions on Affective Computing* **4**, 245–259.

Sehgal, S., Sharma, H. & Anand, A. (2021), 'Smart and context-aware system employing emotions recognition'.

Shackman, A. J., Salomons, T. V., Slagter, H. A., Fox, A. S., Winter, J. J. & Davidson, R. J. (2011), 'The integration of negative affect, pain and cognitive control in the cingulate cortex'.

Shuman, V., Schlegel, K. & Scherer, K. (2015), Geneva emotion wheel rating study properemo view project a developmental perspective of emotion regulation view project, Technical report.
URL: *https://www.researchgate.net/publication/280880848*

ValerioVelardo (2023), 'Generative music ai'.
URL: *https://youtu.be/NpJWprrqlFw?si=OBkGcfxEZROCMmaD*

Vecchiato, G., Maglione, A. G., Cherubino, P., Wasikowska, B., Wawrzyniak, A., Latuszynska, A., Latuszynska, M., Nermend, K., Graziani, I., Leucci, M. R., Trettel, A. & Babiloni, F. (2014), 'Neurophysiological tools to investigate consumer's gender differences during the observation of tv commercials', *Computational and Mathematical Methods in Medicine* **2014**.

Wijethunge, V., Akarawita, S., Hegodaarachchi, T., Abeytunge, S., Gamage, G. & Wickramasinghe, M. (2024), Generative ai and eeg-based music personalization for work stress reduction, *in* 'IECON 2024 - 50th Annual Conference of the IEEE Industrial Electronics Society', pp. 1–6.

Wikipedia (n.d.), 'Melody'.
URL: *https://en.wikipedia.org/wiki/Melody*

Xiong, Z., Wang, W., Yu, J., Lin, Y. & Wang, Z. (2023), 'A comprehensive survey for evaluation methodologies of ai-generated music'.
URL: *http://arxiv.org/abs/2308.13736*

Yasin, S., Hussain, S. A., Aslan, S., Raza, I., Muzammel, M. & Othmani, A. (2021), 'Eeg based major depressive disorder and bipolar disorder detection using neural networks: A review', *Computer Methods and Programs in Biomedicine* **202**, 106007.

Zatorre, R. J., Mori, K., Sang, J. & Wang, X. (2022), 'Emotionbox: A music-element-driven emotional music generation system based on music psychology'.

Zeng, K. (2019), 'A survey on deep learning for multimodal data fusion', *Neurocomputing* **338**, 28–47.

Zhang, W. (2020), Fusion of visual and audio features for emotion recognition in the wild, *in* 'Proceedings of the International Conference on Pattern Recognition', pp. 2270–2277.

Zheng, W.-L. (2015), 'A novel hybrid feature selection method for eeg-based emotion recognition', *Expert Systems with Applications* **42**, 185–195.