ABSTRACT

Music has long been recognized as a powerful medium that influences human emotions, cognition, and behavior. It can energize, calm, inspire creativity, or evoke deep emotional responses. Because of this, music has become an intriguing subject of scientific study, especially in understanding how the human brain perceives and processes different melodies, rhythms, and harmonies.

To better appreciate the impacts that music makes on human brains and which tunes allow having certain feelings such as joy or even disappointment as well as compassion or humor, the patients' brains are observed using EEG. An EEG is a standard and non-invasive procedure, which involves the placement of small leaves, called electrodes, to the skin of the head to measure brain waves. It is also possible to investigate if and what kind of feeling associated with music is always present, and what is the brain that reacts this way to certain music. This is significant because it bears practical consequences regarding the affective states of patients, as well concerning the application of music in the treatment of mental illnesses.

Unfortunately, it is true that for many people, music does in fact elicit strong emotions, but the central processing and neural systems that are responsible for this emotional impact remain mostly a mystery. The EEG system is one of the best in recording neuronal activity during a musical piece. But even today, the area of recording these emotions and the brain responses to music and its emotional fluctuations employing the EEG instrument is still untapped. This research seeks to determine how stimulation through certain tunes is enough to change emotions, and how strong that modification will be.

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

In today's quick and exhausting world, people are almost always looking for suitable ways to relax and express themselves. Music is a universal language that transcends linguistic barriers. It affects emotions, builds creativity, and nurtures connection among individuals. Understanding how and why music has such a profound impact on the human mind has become extremely important in recent years, as it offers valuable insights into both the science of perception and the art of expression.

Music is an eternal form of art. It has the potential to uplift anyone's mind. Firstly, it is necessary to understand what exactly music does to the brain, and how it impacts and affects it. A vast multitude of singular feelings are commonly associated with music, like happiness, sorrow, fear, tension and sadness. We can make use of electroencephalography (EEG) to detect the same. EEG is a non-invasive technique, in which the electrical activity in the brain is recorded, by attaching electrodes to the head. Therefore, by examining the patient's sensitivities to various EEG electrodes, it can help understand how people's feelings change, and thus how music can best be applied.

Chapter 2
PROBLEM STATEMENT

Music is an emotional and psychological phenomenon that can touch anyone at any age, including children. It has a nice ability to improve emotional responses. Many times, it can even handle personal experiences and memories. Many people often connect to music, with happiness, sadness, love, or even excitement. This makes it a shared tool for emotion analysis. Despite this recognized impact, the underlying brain mechanisms to make music so emotionally engaging continue to remain largely unexplored.

Using EEG is considered to be a very impactful research avenue to explore this phenomenon. It records electrical activity in the brain with less invasive methods and it also shows dynamic patterns related to stimuli. This provides an important understanding into the process, using which the brain decodes music. Then it relates it to the emotional response and helps to figure out patterns of activity in the brain.

However, the area of interaction between EEG and music-induced changes in emotions is still unexplored. Although there is knowledge about how music influences mood and behavior, there is not much known about the neural responses of the brain when emotions change following the change in musical tones, tempo, and style.

The research done here aims to bridge the gap present by utilizing EEG to check and find out the emotions evoked by music and their intensity. By scanning the brain activity in multiple channels of EEG, this study will work to give a clearer image of the neural process that goes into

music-evoked emotional states. Finally, it could form a basis of understanding how music might be used as an emotional regulation tool and a mental health support tool.

Chapter 3
LITERATURE REVIEW

3.1. "A EEG-based emotion recognition model with rhythm and time characteristics"

3.1.1. Dataset:

In this case, the DEAP dataset was utilized to investigate the classification of valence and arousal through the EEG signals. This data is of great value in interpreting how these emotional states relate to the characteristic EEG features.

3.1.2. Methodology:

RT-ERM is a model developed for the real-time classification of emotions based on the transmission of EEG signals. The strong point here is the model's processing and responding speed when used in a real life setting.

3.1.3. Performance of existing methods:

RT-ERM reached 62.1 percent in the classification of emotion valence and 69.1 percent in the classification of arousal. These results point out the need for further improvements in order to improve the accuracy level.

3.1.4. Future Scope:

Increased time-scale analysis should help in fostering the stronger understanding of persistent emotional dynamics over longer ranges. This will also enable a better appreciation of the EEG rhythmic pattern.

3.2 "EEG Emotion Recognition Applied to the Effect Analysis of Music on Emotion Changes in Psychological Healthcare"

3.2.1. Dataset:

The DEAP and DREAMER datasets have been adopted which provide a variety of data useful for emotions classification tasks. DEAP concentrates on both dimensions of valence quite the difference that is the arousal while DREAMER has audio and visual stimuli to context the possible emotions.

3.2.2. Methodology:

EER and CNN were applied to these datasets as a feature extraction and recognition approach for emotions classification. Here, CNN is good at extracting spatial patterns while EER is applied in trying to enhance the recognition performance.

3.2.3. Performance of Existing Methods:

DEAP reached 96.4% accuracy, whereas DREAMER reached accuracy of 82.4% demonstrating efficiency of the models constructed. Nonetheless, their performance is however different because of variations in datasets and inputs.

3.2.4. Future Scope:

For better model generalization, collecting emotion on the go can increase data sets. This method would also allow for in the field use.

3.3. "Brain Emotion Perception Inspired EEG Emotion Recognition Using Deep Reinforcement Learning"

3.3.1. Dataset:

EEG signals were utilized for facial emotion classification in SEED and DREAMER datasets. These datasets present different emotional stimuli which allows for an exhaustive assessment of the model's capabilities.

3.3.2. Methodology:

The paper describes in detail the working of a deep reinforcement learning model that aids in showing recognition of emotions in individuals, which is based on the biological mechanisms of the human brain. The focus of the approach is on improving the accuracy of the task of emotion classification by being adaptive to dynamic patterns of the EEG signal.

3.3.3. Performance of Existing Methods:

From the SEED dataset they achieved 91.3% and 88.9% with the DREAMER dataset which all are considerable improvements on previous attempts that used classification based on EEG signals.

3.3.4. Future Scope:

From the above image, the next stage of hope consists of the processes designed for real-time implementations and elaboration of cross-subject generalization employing transfer learning approaches.

3.4 “Music emotion recognition based on temporal convolutional attention network using EEG”

3.4.1. Dataset:

The dataset DEAP included a wide range of data that were necessary for training and validating the model. Additionally, the SWU-M dataset added robustness to the analysis by including different EEG alterations.

3.4.2. Methodology:

CNN-SA-BiLSTM utilizes convolution for the extraction of spatial features, self-attention for the focusing of critical inputs, and bidirectional LSTM for sequence learning. This composite model is very effective at extracting features in both time and space domains.

3.4.3. Performance of Existing Methods:

Accuracies of 92.9% (valence) and 93.17% (arousal) have been obtained thus demonstrating the model's efficient performance in binary emotion classification. These findings are instrumental in the validation of the model's robustness.

3.4.4. Future Scope:

Innovative multi-class emotion classification techniques would help increase the applicability of the model. Furthermore, it provides a way to understand more advanced emotional states.

3.5. "MEEG and AT-DGNN: Improving EEG Emotion Recognition with Music Introducing and Graph-based Learning"

3.5.1. Dataset:

The MEEG data was used for the study of emotion detection through graph-based learning techniques. Its wide variety of data reflects the ability to combine graphs with EEG signals.

3.5.2. Methodology:

The AT-DGNN model uses attention mechanisms and dynamic graphs to capture EEG temporal and spatial dependencies. This novel approach has the capability of modeling complicated relations in EEG data.

3.5.3. Performance of Existing Methods:

Accuracy scores of 83.74% (arousal) and 86.01% (valence) were obtained, which are the possible ways of graph-based learning. The outcomes suggest the efficiency of this novel technique.

3.5.4. Future Scope:

The combination of graph-based models with other architectures would give rise to better performance. Furthermore, this accuracy can be improved for datasets with different features.

3.6. “Recognizing Emotions Evoked by Music Using CNN-LSTM Networks on EEG Signals”

3.6.1. Dataset:

An experimental dataset was used which included EEG recordings from a controlled experimental setting. This dataset was limited but targeted towards positive, negative and neutral emotion data.

3.6.2. Methodology:

Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks are combined in a sequential manner with each layer specializing on a different task – CNN for image feature extracting and LSTM for time series data modeling. This model type is compatible with the analysis of temporal EEG data from the study subjects.

3.6.3. Performance of Existing Methods:

Accuracies of 97.42% and 95.23% for positive negative and for positive negative neutral were obtained. The model performed well in general classification type tasks but not so good in emotion specific tasks.

3.6.4. Future Scope:

More emotion categories could be achieved with the use of a larger emotion dataset. This way, generalization and accuracy of the model would be increased.

3.7 "EEG Emotion Recognition Applied to the Effect Analysis of Music on Emotion Changes in Psychological Healthcare"

3.7.1. Dataset:

For mapping the graphic representation of specific emotions with EEG, the DEAP dataset, which includes compactly structured EEG data, was employed. Such a structured form is best suited for testing technology for the classification of emotions.

3.7.2 Methodology:

The MEC model concentrates on the multi-emotion recognition and classification problem. It encompasses a variety of techniques aimed at improving accuracy and dealing with different dynamics of the recorded EEG.

3.7.3. Performance of the Existing Methods / Techniques:

80% recognition accuracy and 90% classification accuracy have been achieved by the model. These results are indicative of both detection and deep classification of emotions.

3.7.4. Future Scope:

Employing a wider range of styles of music as eliciting material may help account for inter as well as intra-cultural differences in emotional reactions. It might also help improve the accuracy of the identification of certain emotions.

Chapter 4

PROJECT REQUIREMENTS SPECIFICATION

1. Software Requirements

- **Python Environment:** The entire project is implemented in Python, due to its extensive support for machine learning and signal processing libraries.
- **Libraries and Frameworks Used:**
    - **MNE:** Used for loading, visualizing, and preprocessing EEG data, including filtering, artifact removal, and event segmentation.
    - **SpotDL:** Utilized to automatically download songs and metadata directly from Spotify, ensuring consistent audio sources for analysis.
    - **PyDub:** Employed for audio manipulation tasks such as trimming, concatenation, conversion between formats, and waveform analysis.
    - **FFmpeg:** Serves as a backend for PyDub and SpotDL to handle multimedia conversions and efficient audio processing.
- **Deep Learning Models:**
    - **EEGNet:** A compact convolutional neural network (CNN) architecture tailored for EEG signal classification, used here to extract spatiotemporal EEG features relevant to emotional states.
    - **YAMNet:** A pre-trained deep learning model developed by Google, used for extracting audio embeddings and sound event features from songs, which are later mapped to emotion categories.

2. Hardware Requirements

Laptop Specifications:

The proposed system was implemented and tested on a high-performance laptop with the following specifications:

Component	Specification
Processor	Intel® Core™ Ultra 9 185H @ 2.30 GHz
RAM	16 GB (5600 MT/s)
Storage	954 GB SSD
Graphics Card	NVIDIA® GeForce RTX 4060 Laptop GPU (8 GB Dedicated VRAM)
System Type	64-bit Operating System, x64-based Processor

Table 1. Device Specifications

System Justification:

This configuration provided sufficient computational power for deep learning experimentation and model training. The NVIDIA RTX 4060 GPU enabled efficient parallel processing and acceleration of neural network computations, while the Intel Ultra 9 processor and high-speed SSD ensured fast data loading and model checkpointing. Overall, the setup offered an optimal balance of performance and efficiency for running EEG and audio-based multimodal emotion recognition models.

Collaborative Environments used:

Google Colab (T4-GPU)

Chapter 5

SYSTEM DESIGN

Old architecture diagrams:

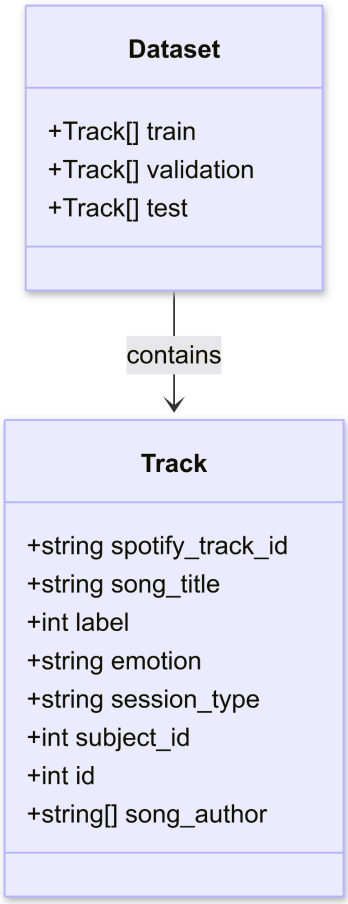

Figure 1: Class Diagram depicting the Metadata

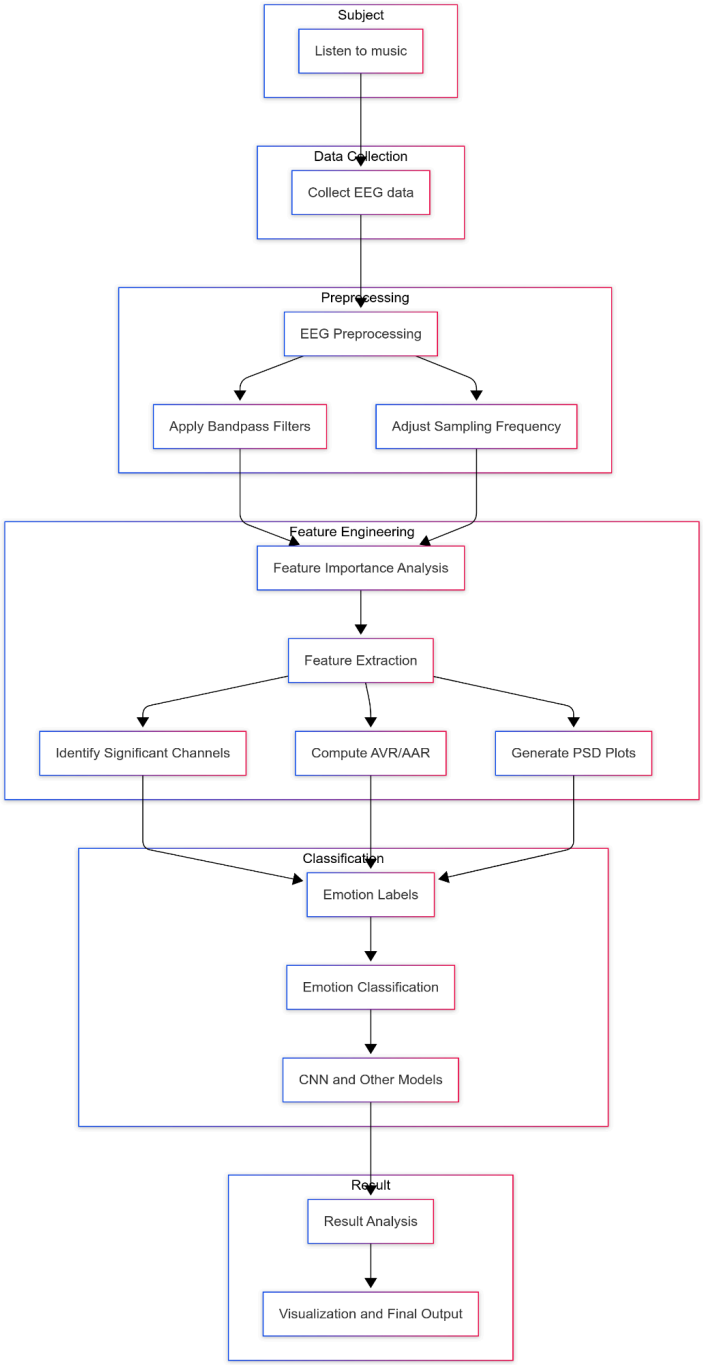

Figure 2: Activity Diagram depicting the Project Workflow

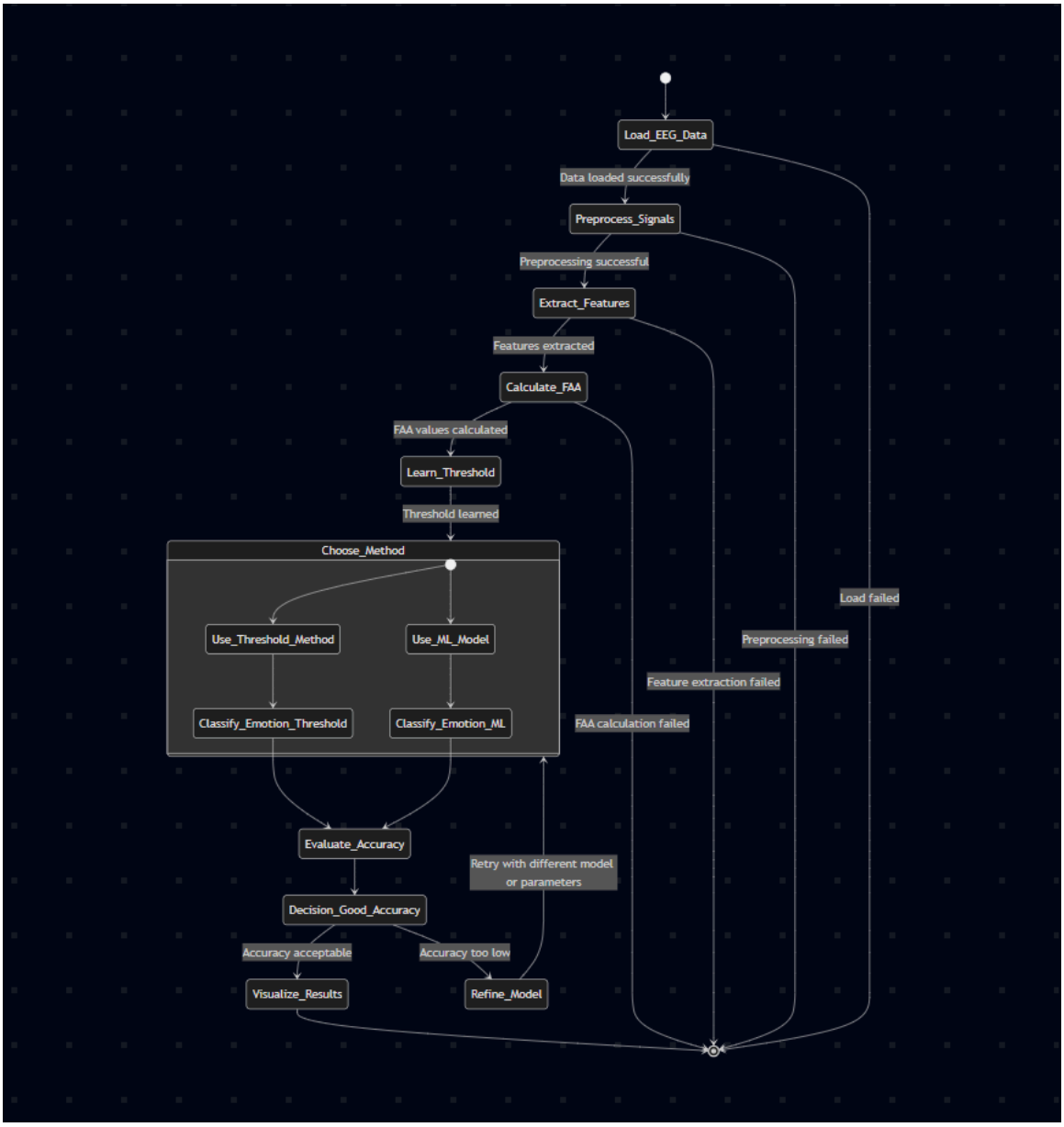

Figure 3: State Diagram

New architecture diagrams:

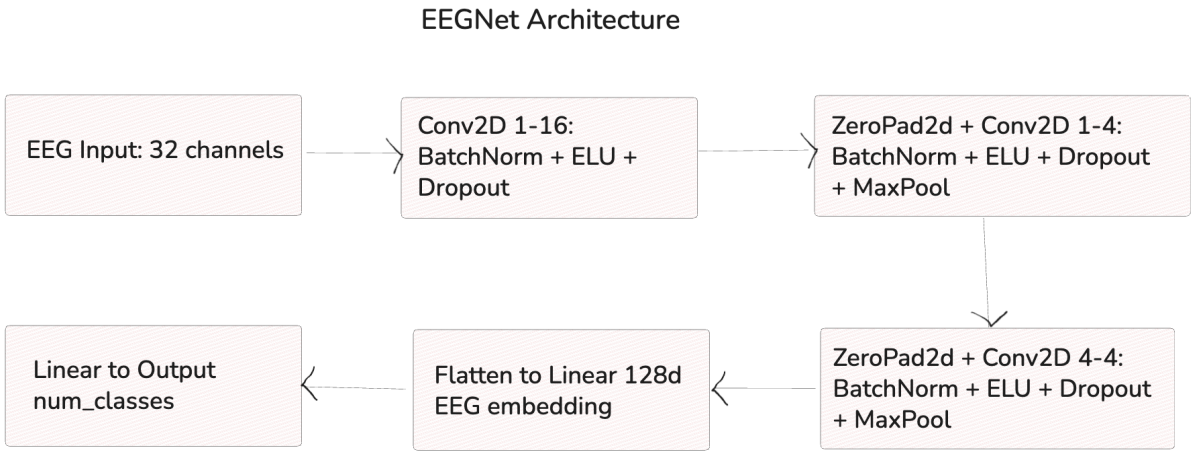

Figure 4: EEGNet Architecture

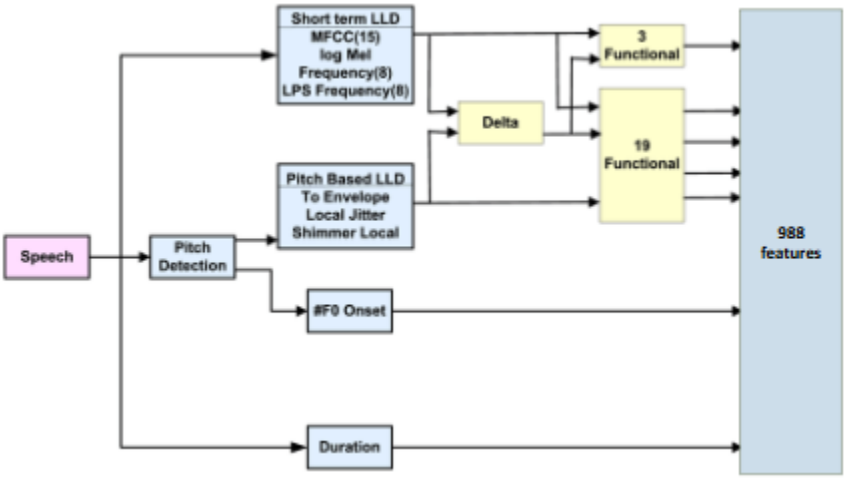

Figure 5: OpenSMILE Architecture

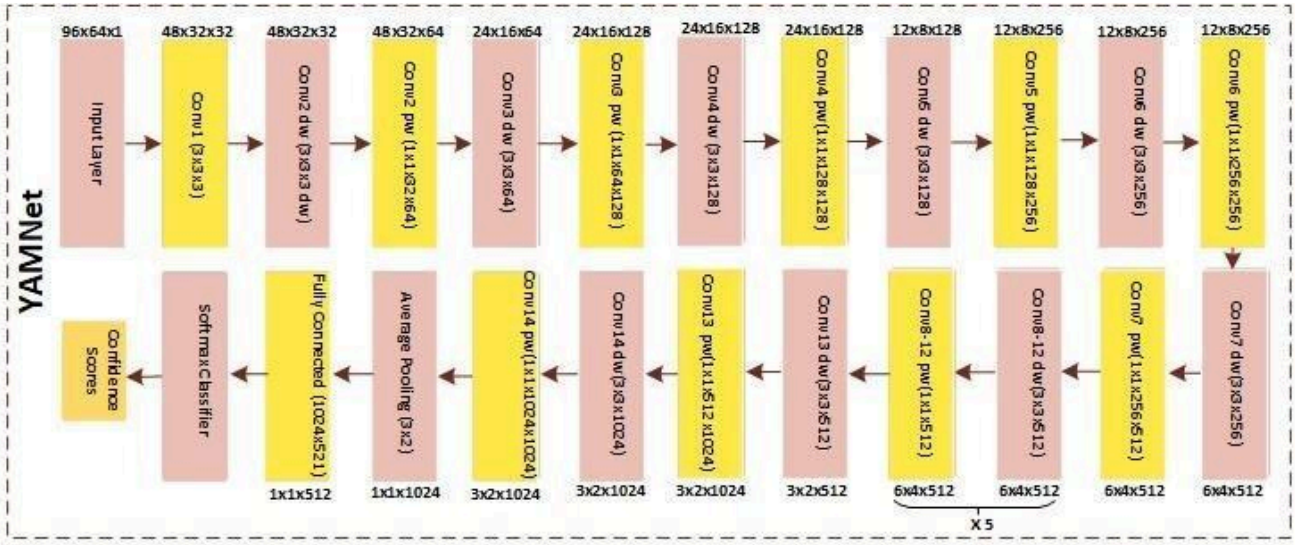

Figure 6: YAMNet Architecture

Chapter 6

PROPOSED METHODOLOGY

6.1. Exploratory Data Analysis (EDA):

6.1.1. Data Preprocessing:

- 78 channels were used to record the EEG data in the original dataset. Some are not required for our analysis and can be considered as noise. We used Independent Component Analysis (ICA) to eliminate muscle and eye movements and other artefacts to choose the 32 channels that were required.
- In order to keep only the necessary frequency bands, we used a Butterworth bandpass filter. This helps us to focus on the particular bands relevant to emotions in the brain.
- We used the standard sampling rate of 128 Hz in order to analyse the EEG data.
- The EEG data is saved in.fif format. We have also converted the data into NumPy array (.npy) format to make it more easily loadable and processable in Python. Such a numerical representation of the data is useful while using the various Machine Learning models for emotion recognition.

6.1.2. Dataset overview:

- Before we begin our implementation, to get an overview of the dataset, we check for any class imbalance for the 21 classes of emotions and compare it with a 2 way, 4 way and 8 way classification by grouping the semantically similar emotions together.

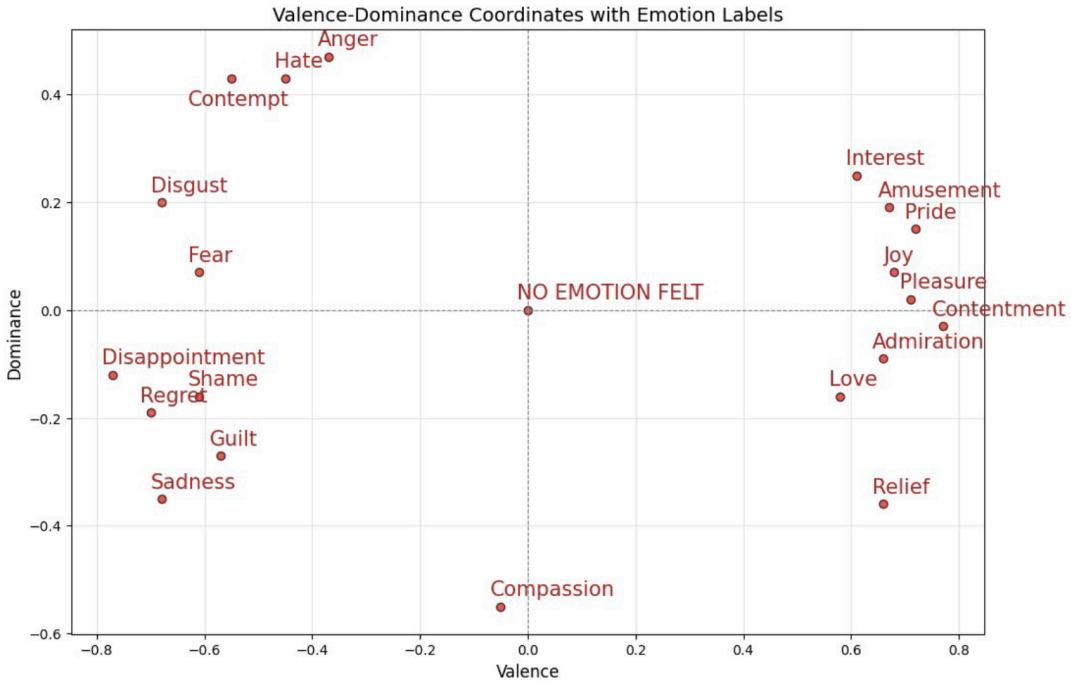

Figure 8: Valence-Dominance coordinate system with emotion labels

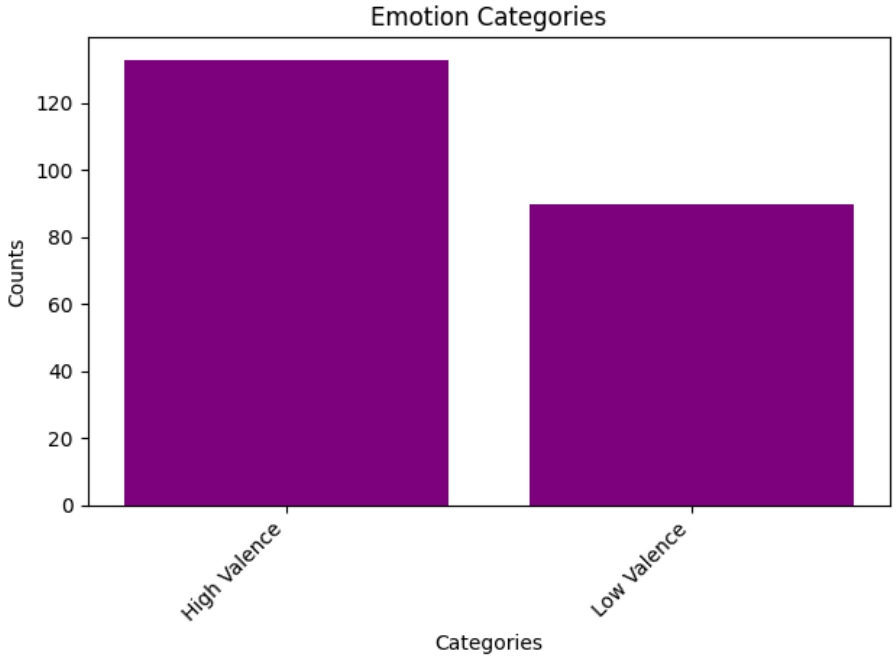

Figure 8: Emotion Counts Histogram for 2 Classes

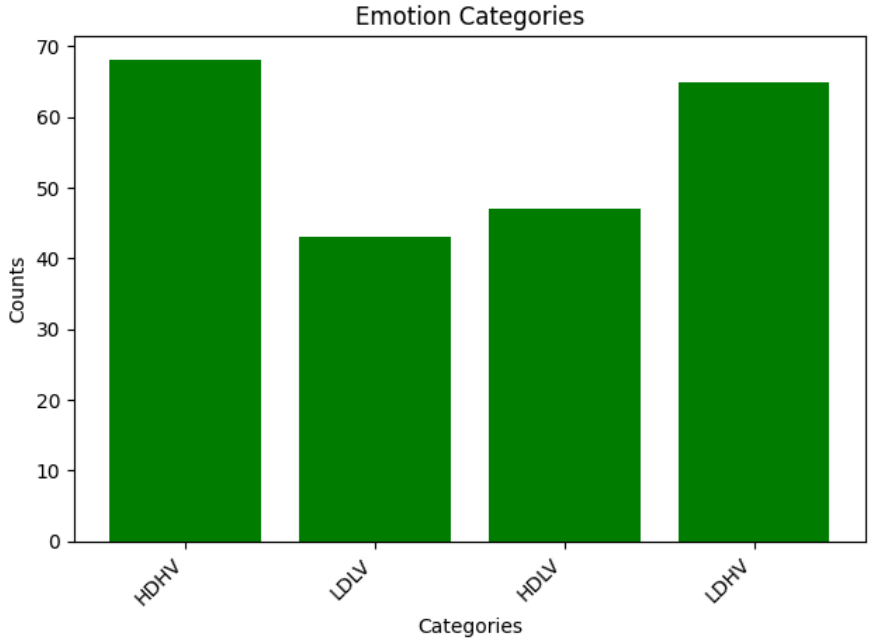

Figure 9: Emotion Counts Histogram for 4 Classes

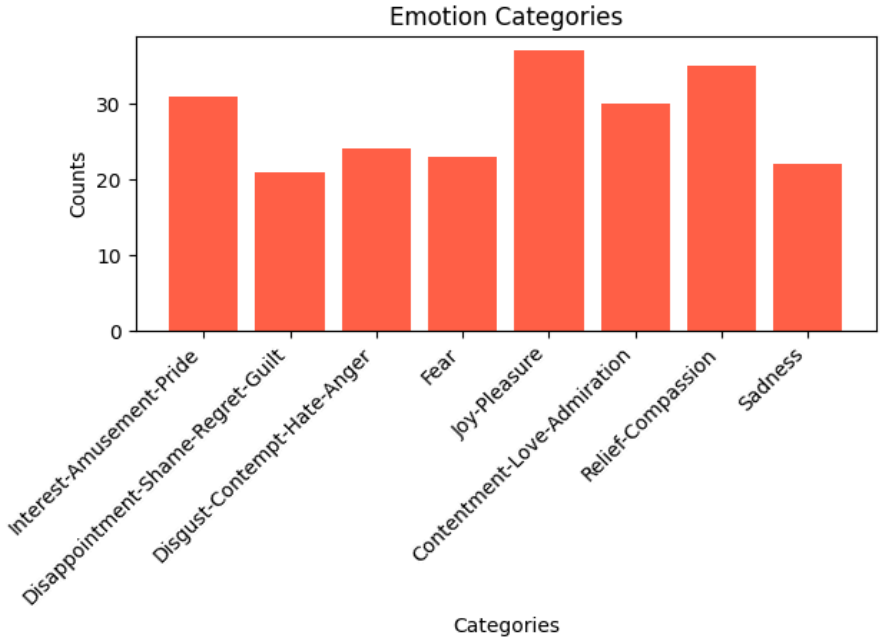

Figure 10: Emotion Counts Histogram for 8 Classes

Categories	Emotions
HDHV	Interest, Amusement, Pride, Joy, Pleasure
LDHV	Contentment, Love, Admiration, Relief, Compassion
LDLV	Sadness, Guilt, Regret, Shame, Disappointment
HDLV	Fear, Disgust, Contempt, Hate, Anger

Table 2: 4-way emotion categories

Class	Emotions
0	Interest, Amusement, Pride
1	Joy, Pleasure
2	Contentment, Love, Admiration
3	Relief, Compassion
4	Sadness
5	Guilt, Regret, Shame, Disappointment
6	Fear
7	Disgust, Contempt, Hate, Anger

Table 3: 8-way emotion categories

6.1.3 Song feature extraction pipeline:

To complement EEG data, we used audio-based emotional context extracted from songs using the OpenSMILE and YAMNet models.

Pipeline Steps:

1. Dataset Preparation:
    - A CSV file containing *track IDs* and *durations* of songs were used during EEG recordings.
2. Song Retrieval:
    - Songs were downloaded using the SpotDL library via Spotify API.
3. Pre-processing:
    - Each track cropped to 90 seconds (matching EEG recording duration).
    - Converted to .wav format for standardized input.
4. Feature Extraction:
    - Each .wav file was processed through the audio models, generating a high dimensional embedding per frame representing auditory features such as pitch, tempo, timbre, and rhythm.
5. Feature Aggregation:
    - Extracted features are averaged across time frames and saved as a single CSV file per song.
    - All song-specific CSVs concatenated to form a master feature file containing 988 dimensions for OpenSMILE and 1024 dimensions for YAMNet for all audio samples.

This pipeline ensures consistent alignment between EEG signals and the corresponding audio stimuli presented during the experiment.

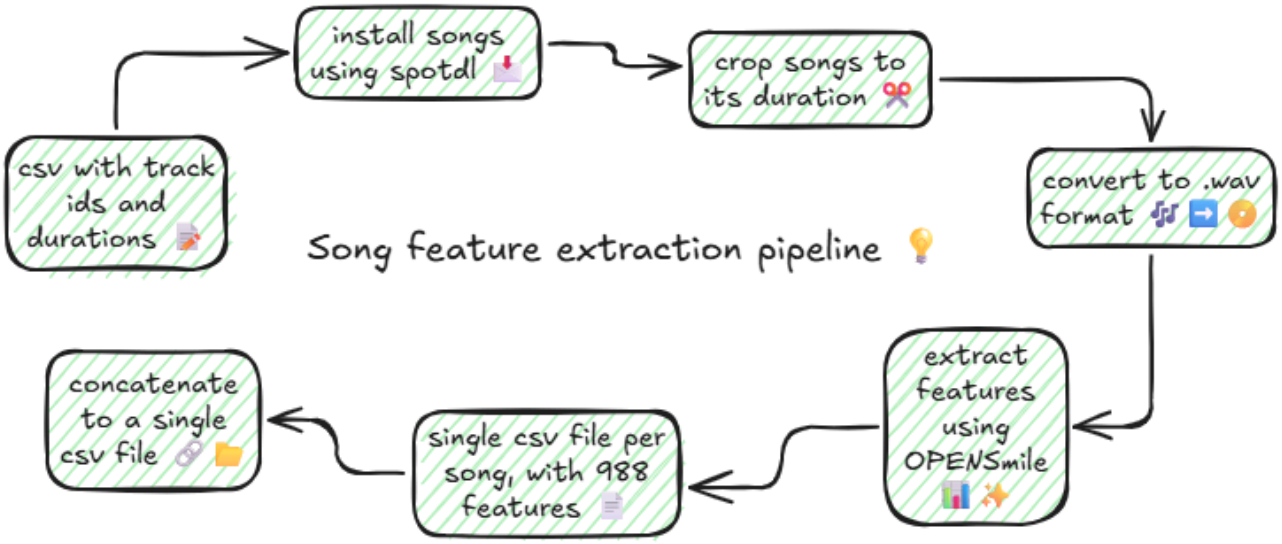

Figure 11: Flowchart representing song feature extraction pipeline for OpenSMILE

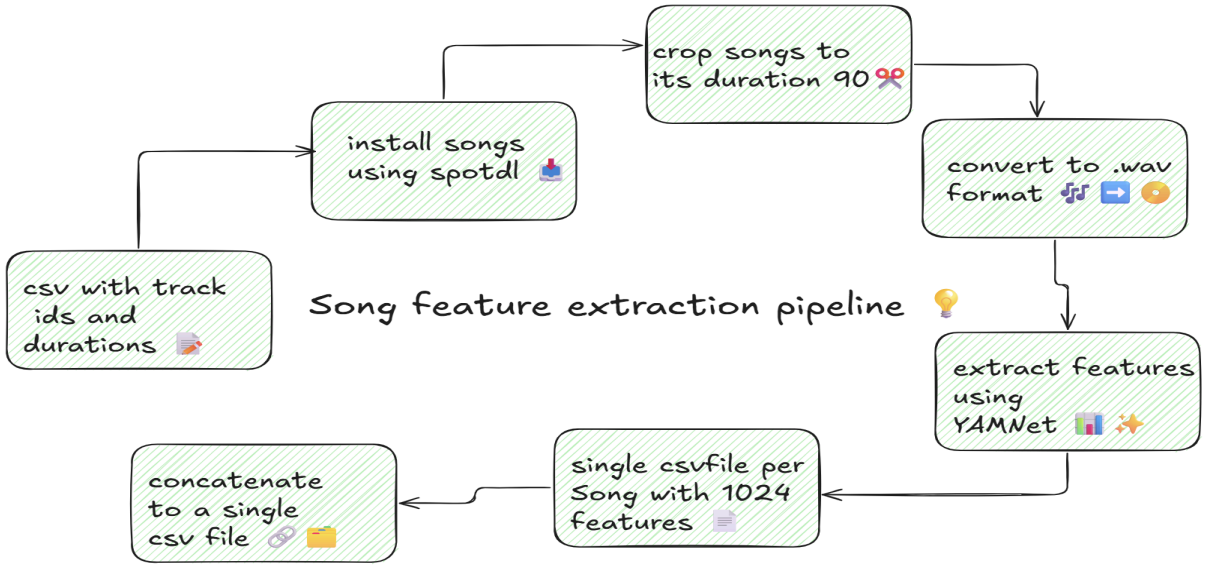

Figure 12: Flowchart representing song feature extraction pipeline for YAMNet

6.2. EEG Feature Extraction:

6.2.1. Asymmetric Variance Ratio (AVR) and Asymmetric Amplitude Ratio (AAR)

We apply Asymmetric Variance Ratio (AVR) and Asymmetric Amplitude Ratio (AAR) for feature extraction. AVR calculates the variance ratio of the signal between the left and right hemispheres. AAR calculates the amplitude (power) ratio between the hemispheres. These are important features because higher values of AVR and AAR correspond to more important channels in the brain, which are useful for emotional classification.

Amplitude Asymmetry Ratio (AAR):

$$AAR = \frac{P(i) - P(j)}{P(i) + P(j)} \quad (2)$$

where $P(i)$ - Spectral power of left hemisphere channel

$P(j)$ - Spectral power of right hemisphere channel

Asymmetric Variance Ratio (AVR):

$$AVR = \frac{V(i) - V(j)}{V(i) + V(j)}$$

where $V(i)$ – Variance of left hemisphere channel

$V(j)$ – Variance of right hemisphere channel

Figure 15: Average Amplitude Ratio (AAR) and Average Variance Ratio (AVR) formulae

6.2.2. Frontal Alpha Asymmetry (FAA)

Frontal Alpha Asymmetry (FAA) is the variation in electrical activity between the right and left frontal lobes of the brain. FAA scores are computed to determine emotional reactions with respect to valence (positive or negative feelings). Positive emotional reactions are measured by higher FAA scores, and negative emotions are measured by lower FAA scores.

$$FAA = \ln\left(\frac{P_{F3}}{P_{F4}}\right)$$

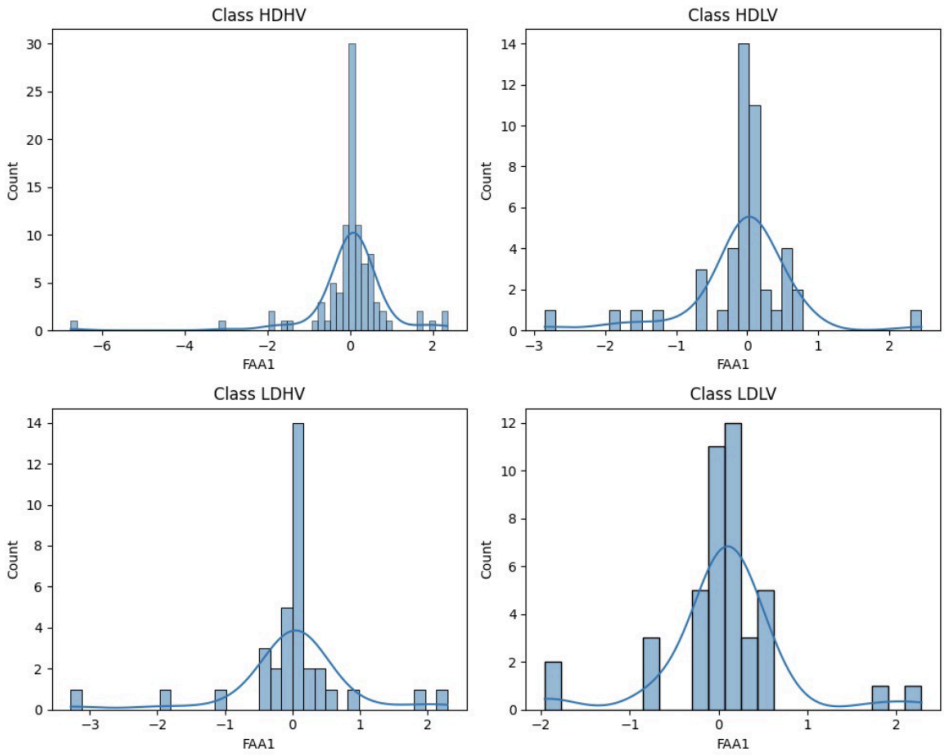

Figure 14: Distribution of Frontal Alpha Asymmetry (FAA) across four emotion classes (HDHV, HDLV, LDHV, and LDLV).

6.2.3. Theta Beta Asymmetry (TBA)

Theta Beta Asymmetry (TBA) reflects the difference between slower theta waves and faster beta waves in the frontal region of the brain. This measure is associated with cognitive effort and several affective processes. Asymmetry at F4-F3 bands is linked to emotional responses, Fp2-Fp1 bands to decision-making, and F8-F7 bands to emotional behaviour. TBA for each electrode pair is calculated as under:

$$\mathrm{TBA}_{\mathrm{pair}} = \ln(P_{\beta}^{\mathrm{Right}}) - \ln(P_{\theta}^{\mathrm{Left}})$$

Here, "P" represents the power band. It is computed as:

$$\hat{P}_{\mathrm{band}}^{c} = \ln\left(\frac{1}{N_{\mathrm{freq}}} \sum_{f_i \in [f_{\min}, f_{\max}]} PSD^{c}(f_i)\right)$$

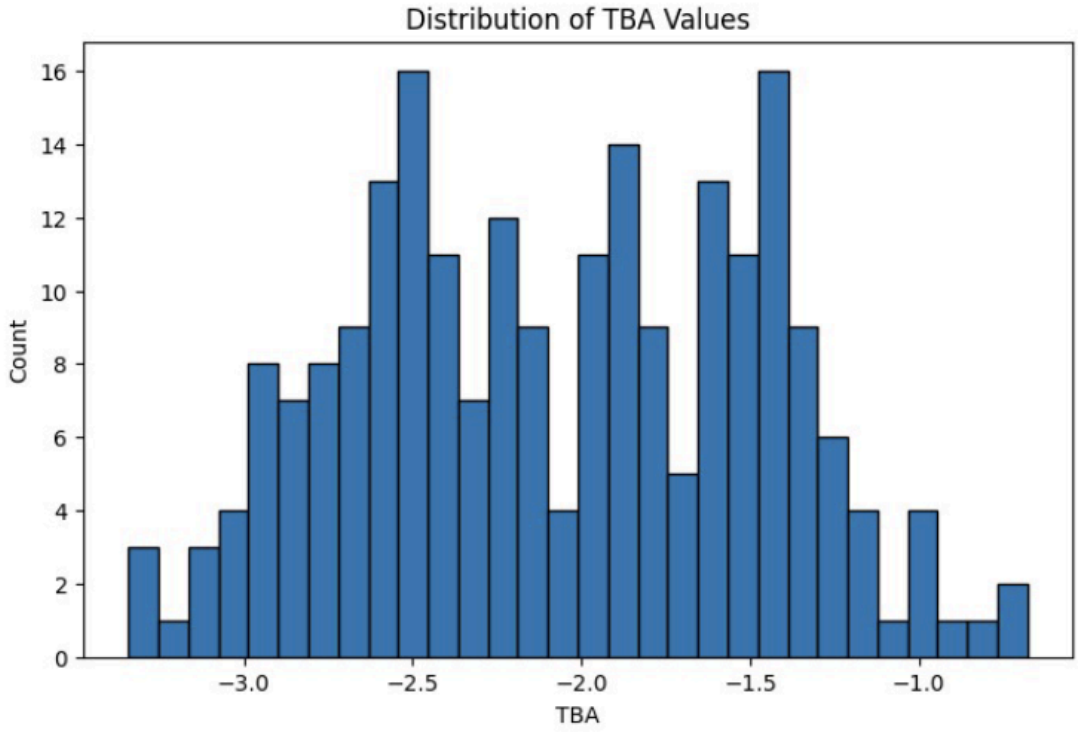

Figure 14: Distribution of Theta–Beta Asymmetry (TBA) values across all EEG samples

- **Threshold Learning for Emotion Classification**

A threshold for FAA values is learned through statistical measures, e.g., Joden's statistic. This is improved with 5-fold cross-validation to get a stable threshold value. The learned threshold distinguishes between positive and negative emotions based on the FAA values.

6.3. Data Visualization

6.3.1. AAR and AVR plots

We visualize AVR and AAR features to determine the importance of various brain regions in emotional processing. This is done by plotting it on a montage plot according to electrode placement on the skull. Power Spectral Density (PSD) displays the frequency distribution of the electrical activity of the brain and helps in identifying specific emotional responses.

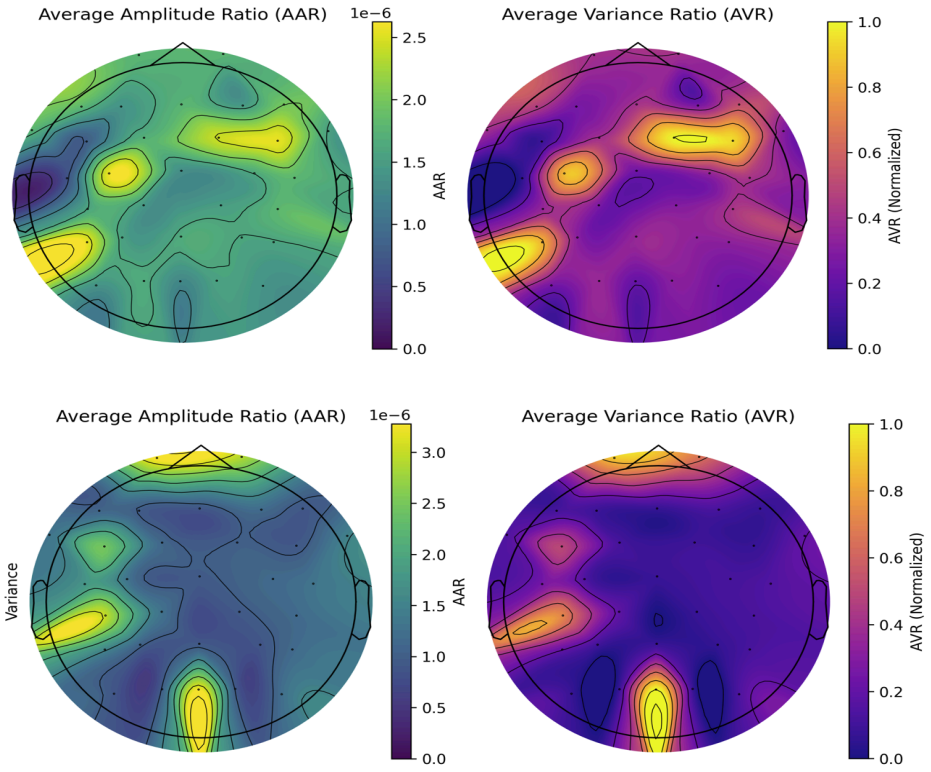

Figure 14: Average Amplitude Ratio (AAR) and Average Variance Ratio (AVR) plots

- **Observations**

From the exploratory analysis, we can see that the most important EEG channel pairs are found in the middle of the skull. These areas of the brain help to notice the changes in the brain activity and the emotional responses, especially for music perception.

6.3.2. Polar plots:

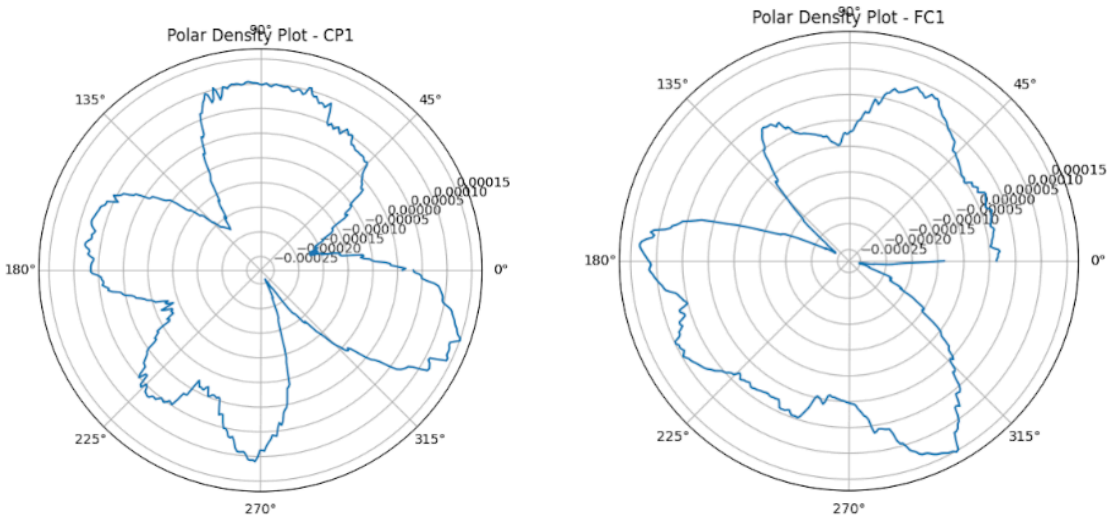

Figure 16: Polar plots for significant and non significant electrodes

Polar plots are graphs that represent EEG data in terms of a radius and an angle, making it extremely useful in EEG-based emotion recognition. In general, it is noticed that butterfly shaped spatial plots for a particular electrode (out of the 32), is a direct indication of that electrodes significance in predicting a particular emotion. Conversely, electrodes that don't have a typical butterfly shape are considered to be insignificant This can be explained quite simply by attributing the increased significance to the fact that the overlapping multichannel dynamics carry more discriminative information than angle-based radial mappings. Although this seems to be a useful metric, it is difficult to generalize a common set

of electrodes for individuals and emotions. It varies largely based on music preferences and emotions felt while listening to them. Though this was a worthwhile exploration, giving us a deeper understanding of how electrodes affect emotions, we couldn't move ahead with it due to the lack of generalization.

6.3.3. Power Spectral Density (PSD)

The dataset combined EEG-based PSD representations with acoustic features extracted from songs to enable classification based on valence. Each EEG recording was converted into a 2D PSD image. These images captured spatial–spectral power distributions across channels, while corresponding song-level features (e.g., tempo, spectral contrast, timbre statistics) were incorporated to enrich contextual representation. All PSD images were precomputed and standardized to a fixed spatial resolution before being transformed into tensor inputs. Numerical song features were normalized using z-score scaling and aligned with EEG samples through unique recording identifiers to ensure feature correspondence.

A hybrid deep learning model integrating convolutional and fully connected components was used. The EEG branch consisted of a 2D-CNN that extracted hierarchical spatial–frequency features from PSD images These used ReLU activations and max-pooling operations for translation-invariant feature extraction. Additionally, a fully connected subnetwork processed the acoustic features to generate compact latent embeddings. These two feature spaces were concatenated and passed through dense layers with Softmax activation to predict valence. Model training was performed using the Adam optimizer with categorical cross-entropy as the loss function. To assess generalization performance and reduce overfitting, a five-fold cross-validation strategy was adopted. As part of this, the model was independently trained and validated across multiple data splits. The training and validation curves were also visualized to determine stability across folds.

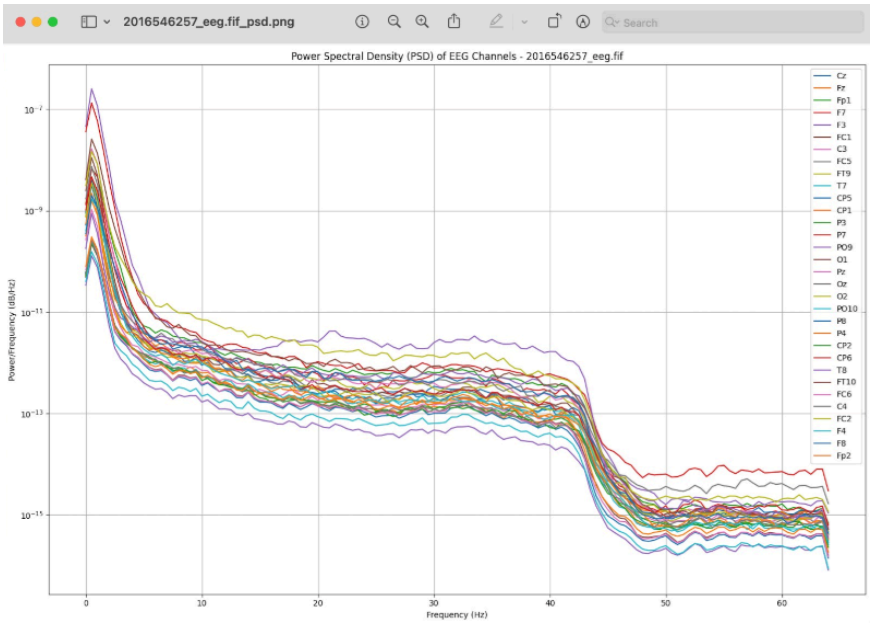

Figure 13: Power Spectral Density Plot

6.4. Machine Learning Classification

After the threshold has been set machine learning algorithms are used for the emotion classification. Algorithms utilized include Logistic Regression, SVM, Random Forest. The models assist in classifying EEG data depending on the FAA value in comparison to the learned threshold.

6.4.1. Classification and Accuracy

With the FAA values calculated, if the FAA of a specific entry is over the threshold learned, it is labeled as a positive emotion. If it is under the threshold, it is labeled as a negative emotion. The correctness of the labels is determined by comparing the predicted labels to the labels in the given dataset.

6.4.2. Review

In Phase 2, the focus was primarily on building the baseline EEGNet model and establishing the data preprocessing pipeline for EEG and audio signals. While the initial

architecture produced encouraging results, it showed limitations in capturing deeper temporal–spatial dependencies and in handling multimodal inputs effectively.

Phase 3 builds upon this foundation by refining the model architecture, introducing multi-scale feature learning, integrating song-based embeddings, and optimizing the training pipeline for improved accuracy and generalization. The enhanced model, termed GESSNet (Global EEG Spectral Spatial Network), extends EEGNet with adaptive and multi-scale design principles while enabling multimodal fusion between EEG and audio features.

The model consists of two parallel branches — one for EEG feature learning and the other for audio feature extraction, followed by a fusion and classification module.

The modified architecture builds upon EEGNet by integrating additional convolutional and dense layers to enhance the model's ability to capture complex spatio-temporal features from EEG signal

Chapter 7

IMPLEMENTATION AND PSEUDO CODE

7.1. Baseline EEGNet Architecture

The base model for EEG signal analysis is adapted from EEGNet, a compact and efficient convolutional neural network specifically designed for EEG-based brain–computer interface (BCI) applications.

The model takes as input EEG signals of 32 channels and processes them through multiple convolutional and normalization layers to generate meaningful spatial and temporal embeddings.

Architecture Flow:

1. EEG Input:
    - Shape: *(32 channels × time samples)*
    - Represents raw EEG recordings filtered and normalized before feeding into the network.
2. Convolutional Block 1:
    - Conv2D (1 → 16 filters) with small temporal kernels to extract fine-grained features.
    - Batch Normalization stabilizes learning.
    - ELU activation (Exponential Linear Unit) introduces non-linearity while maintaining smoother gradients.
    - Dropout applied to prevent overfitting.
3. Convolutional Block 2:
    - Zero-padding ensures consistent output dimensions.
    - Conv2D (1 → 4 filters) followed by BatchNorm, ELU, Dropout, and MaxPooling to reduce dimensionality and extract more abstract spatial patterns.

4. Convolutional Block 3:
    - Zero-padding + Conv2D(4 → 4 filters).
    - Similar operations (BatchNorm + ELU + Dropout + Pooling).
    - This layer captures long-term temporal dependencies.
5. Flattening and Dense Layer:
    - Output features are flattened into a 128-dimensional EEG embedding, representing spatial–temporal brain activity.
    - A fully connected layer projects the embedding to the number of output emotion classes.

7.1.1 Limitations of Original EEGNet

- Shallow architecture fails to learn hierarchical representations.
- Fixed filter sizes do not adapt to different brainwave frequencies (alpha, beta, theta, gamma).
- Lack of global temporal context limits cross-channel information learning.

7.2 Model Architecture Improvements

Although EEGNet provides a lightweight architecture, its shallow depth (three convolutional layers) limits its ability to capture complex spatio-temporal dependencies inherent in EEG signals. To overcome these drawbacks, we developed a modified EEGNet variant – GESSNet, which introduces multi-scale convolutional layers and deeper dense connections. We further augment audio features extracted using OpenSMILE and YAMNet audio models to GESSNet for enhanced accuracy. We also propose a model based on neural features fused with acoustic features from YAMNet - called NeuroMENet.

7.2.1 Modified EEGNet (GESSNet - Global EEG Spatial Spectral Network)

To overcome the limitations of baseline EEGNet variants and tailor the architecture for music-evoked emotion recognition, we designed GESSNet. It brings together:

- **Global features:** Captures long-range temporal dependencies through global pooling operations.
- **Spatial features**: Exploits spatial relationships between EEG channels (channel topology and inter-region dynamics)
- **Spectral features**: Emphasizes frequency-specific dynamics critical in music-based EEG responses.

7.2.2. Key Improvements over baseline EEGNet:

Parameter	EEGNet (Baseline)	Modified EEGNet (GESSNet)
Activation	ReLU	ELU (smoother, prevents dying neurons)
Regularization	Standard dropout	Stronger dropout (0.5)
Normalization	None	Batch Normalization added
Filters	Fixed	Adaptive filters dynamically adjusted per layer
Model Type	Static	Dynamic architecture for EEG variability

Table 4. GESSNet Advantages

These changes improved generalization and convergence speed, especially on EEG datasets with inter-subject variability.

This three-fold combination enhances sensitivity to both global temporal rhythms induced by music and localized spectral-spatial EEG patterns associated with affective processing. The architecture preserves the advantages afforded by EEGNet while introducing multi-scale temporal kernels to capture both fast transient EEG bursts and slower, music-aligned patterns and spatial-spectral

attention block to dynamically reweigh EEG channels, emphasizing frequency–region interactions relevant for emotion recognition. We implemented global pooling across channels and time to improve robustness to the variability between subjects while simultaneously maintaining discriminative temporal features.

Kernel Size	Frequency Band	Feature Captured
[1, 16]	Gamma (high frequency)	Short-term fluctuations
[1, 32]	Beta & Alpha	Medium-term patterns
[1, 64]	Theta	Long-term dependencies

Table 5. GESSNet kernel sizes & feature bands

These branches are concatenated and passed through deeper dense layers, allowing simultaneous extraction of local and global temporal context.

Advantages:

- Multi-scale feature learning across frequency bands.
- Better generalization for varying EEG sampling rates.
- Improved accuracy and robustness against noise.

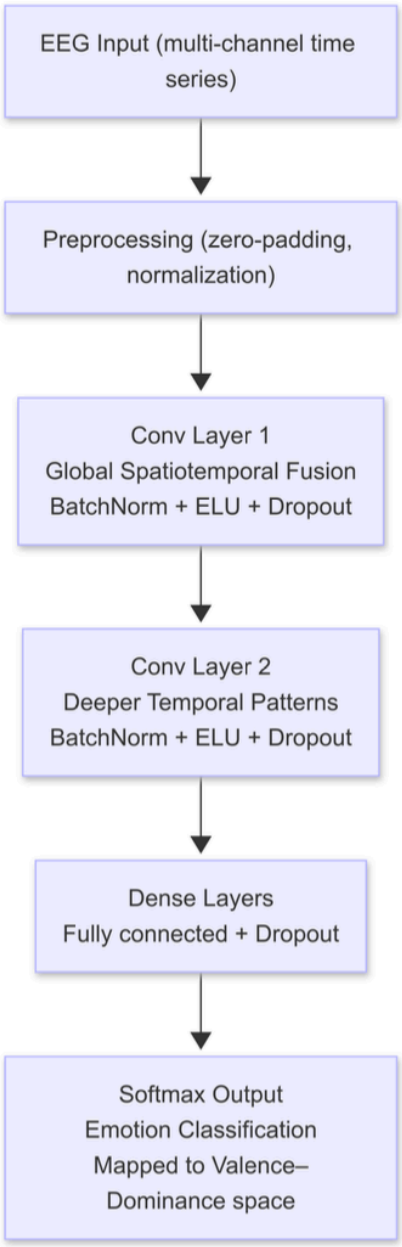

Figure 17: Flowchart illustrating the architecture of GESSNet

7.5 Overall Integrated Architecture - Fusion of EEG and Audio features (OpenSMILE and YAMNet)

The overall model integrates both EEG and song-based features through feature fusion followed by joint emotion classification.

7.5.1 EEG Branch

- Input: 64 x time
- Conv2D (1 → 8) → DepthwiseConv (8 → 16) → AdaptivePooling (512)
- Fully Connected Layer: 512 → 128
- Output: 128-dimensional EEG embedding

7.5.2 Song Feature Branch

- Input: 1024-dimensional song embedding
- Fully Connected Layer 1: 1024 → 256 (ReLU)
- Fully Connected Layer 2: 256 → 32
- Output: 32-dimensional song embedding

7.5.3 Feature Fusion and Classification

- Concatenation of EEG (128 d) and Song (32 d) features → 160-dimensional joint representation
- A fully connected layer maps this to 4 emotion classes (e.g., happy, sad, calm, and excited).
- Softmax layer used for final probability distribution across classes.

This fusion enables the model to jointly interpret physiological (EEG) and contextual (audio) cues, providing a holistic emotion representation.

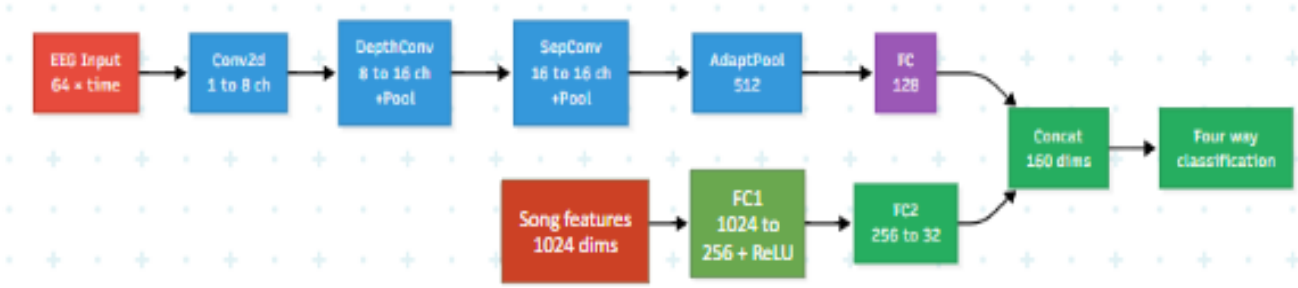

Figure 7. Overall architecture

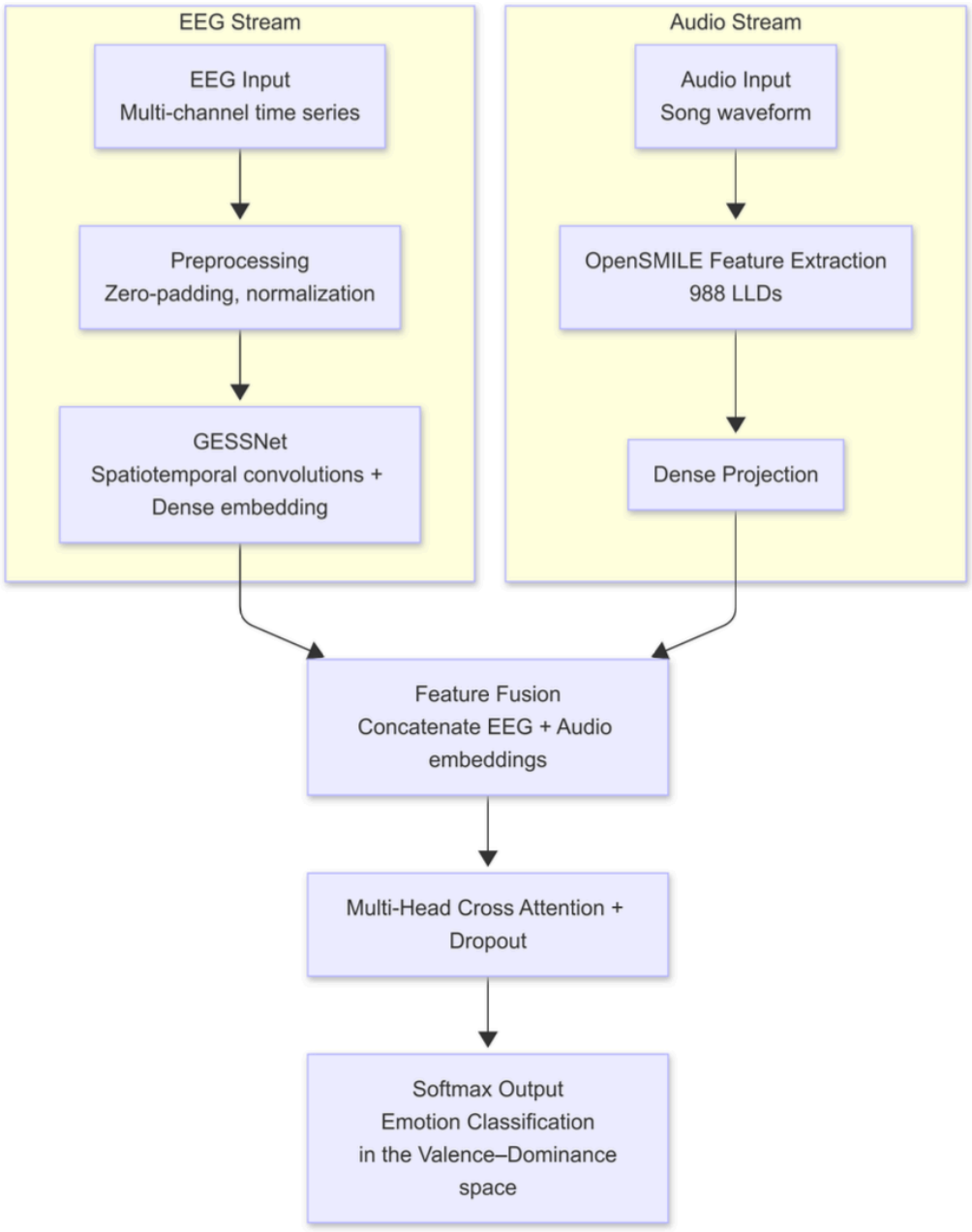

Figure 18: Flowchart illustrating the architecture of GESSNet + OpenSMILE

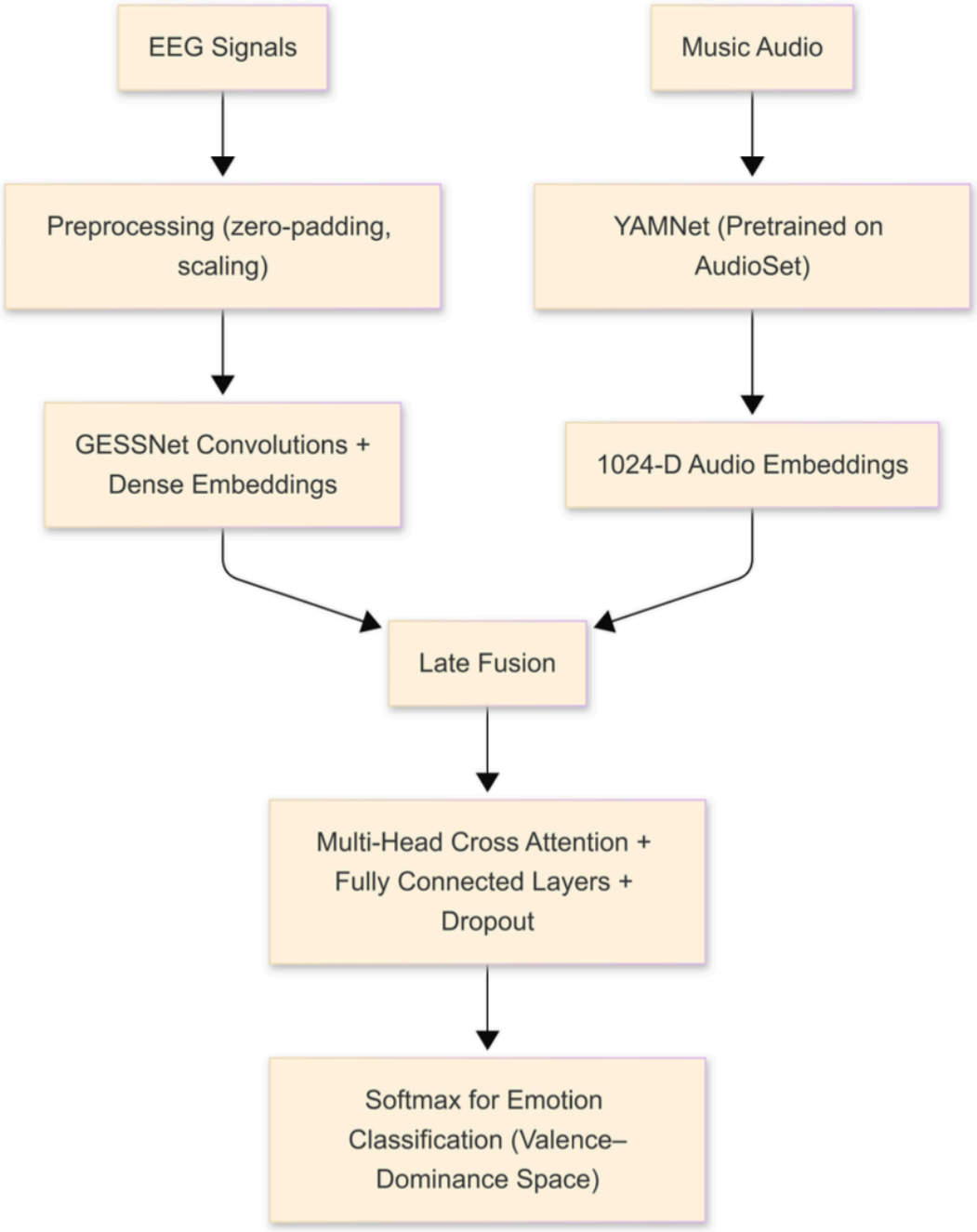

Figure 19: Flowchart illustrating the architecture of GESSNet + YAMNet

7.6. NeuroMENet (Neural Music Emotion Network)

NeuroMENet is a deep learning architecture designed to understand emotions by combining information from brain signals and music attributes. The model takes in EEG asymmetry features (viz. FAA and TBA) along with YAMNet audio embeddings that capture acoustic properties . NeuroMENet is a deep feedforward neural network with five hidden layers containing 1024, 512, 256, 128, and 64 neurons, respectively. Each layer employs ReLU activation to help the model learn complex patterns, while batch normalization and dropout are added to stabilize training and reduce overfitting. Further, to improve generalization, L2 regularization is applied. The network ends with a softmax output layer that classifies emotions into either two or four categories. It is trained using the Adam optimizer and sparse categorical cross-entropy loss, with early stopping mechanism. To address the issue of class imbalance, class weights are assigned using a balanced strategy. Overall, NeuroMENet blends neural and acoustic information in a structured way to recognize emotional patterns more accurately and reliably.

```
Function build_model(input_dim, num_classes)
model ← Sequential()
model.add(Dense(1024, activation='relu', input_shape=(input_dim,), kernel_regularizer=L2(1e-4)))
model ← BatchNormalization()
model ← Dropout(0.5)
model.add(Dense(512, activation='relu', kernel_regularizer=L2(1e-4)))
model ← BatchNormalization()
model ← Dropout(0.4)
model.add(Dense(256, activation='relu', kernel_regularizer=L2(1e-4)))
model ← BatchNormalization()
model ← Dropout(0.3)
model.add(Dense(128, activation='relu', kernel_regularizer=L2(1e-4)))
model ← BatchNormalization()
model ← Dropout(0.2)
model.add(Dense(64, activation='relu', kernel_regularizer=L2(1e-4)))
model ← BatchNormalization()
model ← Dropout(0.1)
model.add(Dense(num_classes, activation='softmax'))
model.compile(optimizer='adam', loss='sparse_categorical_crossentropy', metrics=['accuracy'])
return model
```

Figure 20: Algorithm representing NeuroMENet architecture

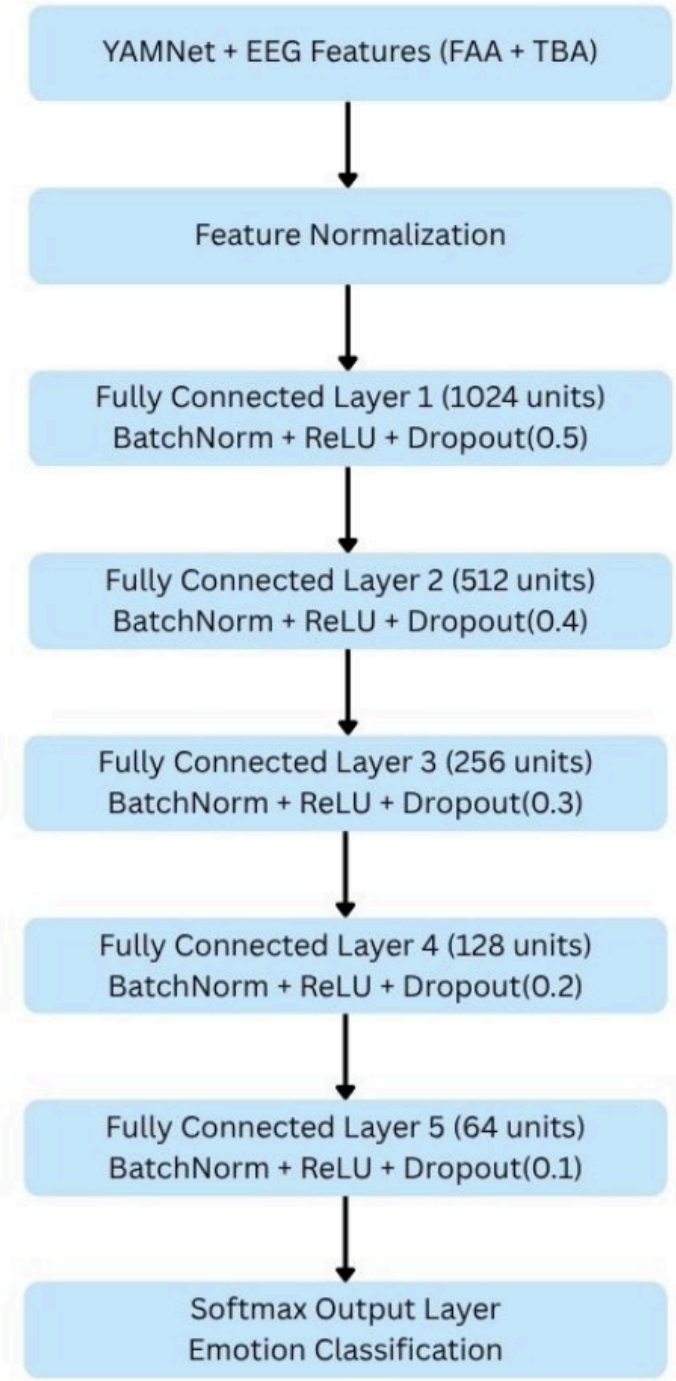

Figure 21: Flowchart illustrating NeuroMENet architecture

7.6 Training and Optimization

- Loss Function: Cross-Entropy Loss
- Optimizer: Adam (learning rate = 0.001)
- Batch Size: 32
- Epochs: 50–100 based on convergence
- Dropout Rate: 0.5 (applied after each major layer)
- Evaluation Metrics: Accuracy, Precision, Recall, and F1-score

Pseudocode:

1. Initialize HYPERPARAMETERS (batch_size, epochs, lr, patience, dropout)

2. LOAD YAMNetfeatures from CSV
   → Normalize using StandardScaler
   → Remove low-variance features
   → Store in dictionary {song_id: feature_vector}

3. FOR each EEG file:
   load EEG data
   pad/center to fixed length
   match corresponding song features
   augment EEG (noise, scaling)
   store (EEG, song_features, label)

4. SPLIT dataset → train, validation sets

5. DEFINE GESSNet+YAMNet Model:
   EEG branch → CNN + SeparableConv layers
   Song branch → Dense + Attention
   Cross-attention → Fuse EEG + song features
   FC layers → Predict emotion class

6. SET loss = Weighted CrossEntropy
   SET optimizer = AdamW
   SET scheduler = CosineAnnealingLR
   INIT early stopping

7. FOR epoch in range(EPOCHS):
   TRAIN:

```
    for (EEG, song, label) in train_loader:
        y_pred = model(EEG, song)
        loss = criterion(y_pred, label)
        backpropagate(loss)
  VALIDATE:
    compute val_loss, val_accuracy
  UPDATE scheduler
  IF val_accuracy improves → save model
  ELSE → increment patience counter
  IF patience exceeded → break
```

Chapter 8
RESULTS AND DISCUSSION

The proposed GESSNet architecture was evaluated for four-class emotion classification, achieving significant improvements over the competition baseline. The results are illustrated as under:

Model Name	Accuracy	Std. Dev
GESSNet	31.3%	± 3.4%
GESSNet + OpenSMILE	45.4%	± 2.9%
GESSNet + YAMNet	47.4%	± 1.5%
NeuroMENet	42.78%	± 1.59%

Table 6. Model Results for 4-way classification across all models

Modality	Accuracy	Std. Dev.
2-way	88.3%	±1.80%
4-way	47.4%	±1.50%
8-way	29.0%	±2.81%

Table 7. Model Results for 2, 4, 8-way classification for GESSNet + YAMNet

The inclusion of YAMNet features contributed to enhanced representation of emotional cues from the audio signals, effectively capturing both spectral and temporal aspects of emotions. The smaller variance observed in the GESSNet + YAMNet setup indicates that the model not only achieved higher accuracy but also demonstrated consistent performance across multiple runs. This shows that

transfer learning from large-scale pretrained audio networks such as YAMNet significantly benefits emotion classification tasks.

The comparison between different model configurations highlights the importance of multimodal feature integration. While GESSNet alone showed promising baseline results, combining it with external acoustic features allowed the model to leverage richer information about pitch, tone, and rhythm — all critical to emotional perception. These findings confirm that hybrid architectures integrating deep and handcrafted features can outperform conventional single-source models in emotion recognition.

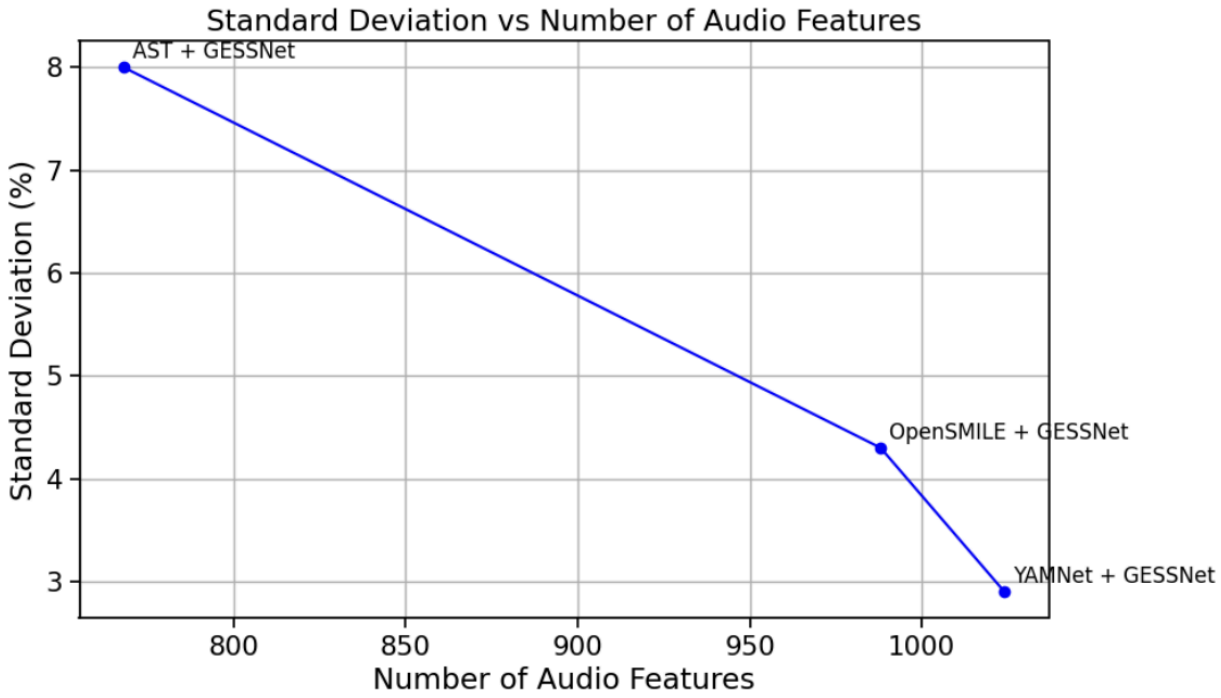

Figure 6. Line graph of Standard Deviation vs. Number of Features in Audio Models

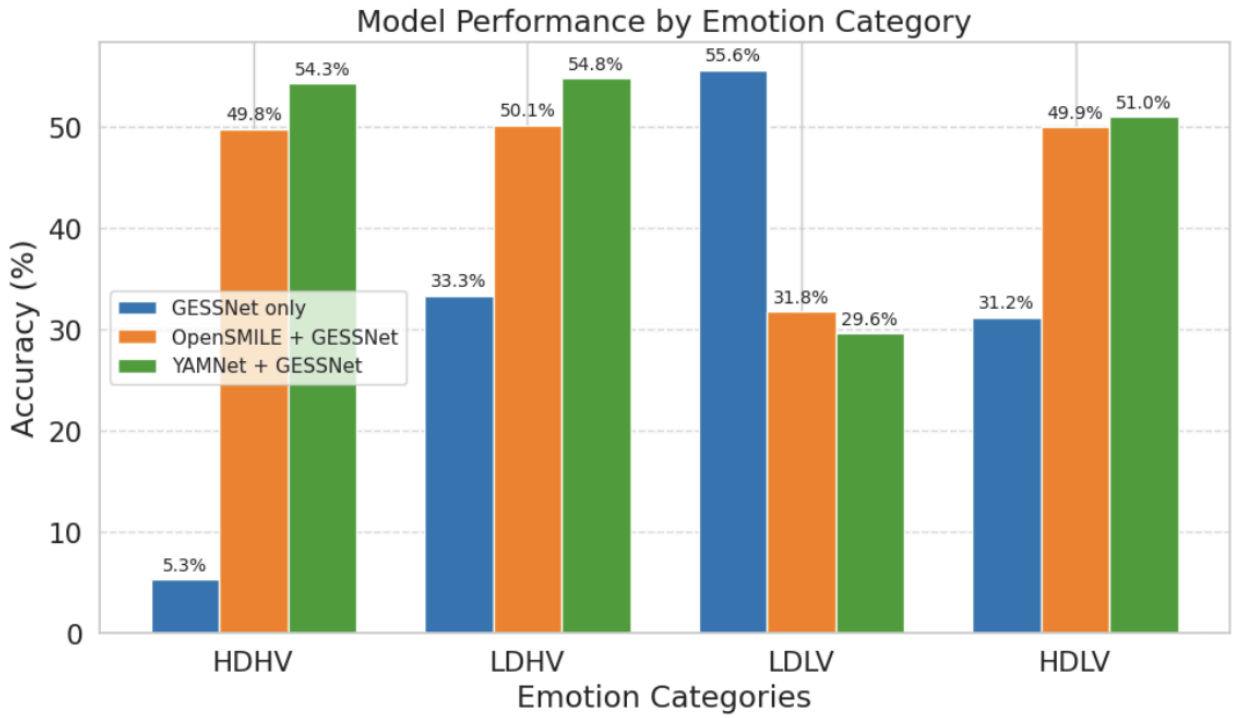

Figure 6. Model performance by Emotion Category

Chapter 9

CONCLUSION AND FUTURE WORK

Through this work, we have successfully developed and demonstrated a novel hybrid deep learning framework, **GESSNet + YAMNet**, capable of classifying emotions into four distinct categories with significant accuracy improvements over baseline models. We also introduce **NeuroMENet**, which is a robust architecture that effectively integrates EEG-based physiological signals and audio-based contextual features, leading to enhanced emotion recognition performance. The combination of adaptive multi-scale convolutions in GESSNet and deep audio embeddings from YAMNet proved to be computationally efficient, enabling the extension of this study to successfully classify emotions into 2, 4 and 8-way as well. Our model performed exceptionally well on all three modalities, effectively surpassing the ICASSP baselines for each.

Beyond improving classification accuracy, the use of asymmetry-based EEG features offers interpretability and computational efficiency. This helps make the framework suitable for real-time or embedded emotion-aware applications. Despite these advances, the observed decline in performance with increasing granularity of emotions highlights ongoing challenges in modeling subtle affective states.

Future work will focus on -

(1) employing transformer-based cross-modal encoders for enhanced temporal alignment

(2) generalizing to other datasets to include diverse musical genres and subject populations to strengthen model generalization.

APPENDIX A:
DEFINITIONS, ACRONYMS AND ABBREVIATIONS

- **EEG:** Electroencephalogram
- **PSD:** Power Spectral Density
- **FAA:** Frontal Alpha Asymmetry
- **TBA:** Theta Beta Asymmetry
- **AVR:** Asymmetric Variance Ratio
- **AAR:** Asymmetric Amplitude Ratio
- **CNN:** Convolutional Neural Network
- **ReLU**: Rectified Linear Unit
- **ELU**: Exponential Linear Unit
- **OpenSMILE: Open** Speech and Music Interpretation by Large scale feature Extraction
- **YAMNet:** Youtube Audio Model (by Google)
- **GESSNet:** Global EEG Spatial Spectral Network
- **NeuroMENet:** Neural Music Emotion Network