Multiclass Emotion Classification by using Spectrogram Image Analysis: A CNN-XGBoost Fusion Approach

by

Ishraq Ahmed Esha
19301261
Shahrin Rahman
20101464
Sayem Kader Chowdhury
19101076
Jannatul Ferdous Mim
23141096

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
School of Data and Sciences
Brac University
May 2023

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Ishraq Ahmed Esha
19301261

Shahrin

Shahrin Rahman
20101464

Sayem

Sayem Kader Chowdhury
19101076

Mim

Jannatul Ferdous Mim
23141096

Approval

The thesis titled "Multiclass Emotion Classification by using Spectrogram Image Analysis: A CNN-XGBoost Fusion Approach" submitted by

1. Ishraq Ahmed Esha(19301261)
2. Shahrin Rahman(20101464)
3. Sayem Kader Chowdhury(19101076)
4. Jannatul Ferdous Mim(23141096)

Of Spring, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on May 29, 2023.

Examining Committee:

Supervisor:
(Member)

Md. Golam Rabiul Alam, PhD
Professor
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

RRahman

Rafeed Rahman
Lecturer
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Chair)

Md. Golam Rabiul Alam, PhD
Professor
Department of Computer Science and Engineering
Brac University

Ethics Statement

This thesis was prepared based on the outcomes of our in-depth investigation, we the authors hereby truly proclaim. The sources that were used are highly acknowledged and credited in this study. This research paper has never been sent to any organization for certification.

Abstract

In recent times, AI based emotion recognition also known as affective-computing which is a burgeoning branch of artificial intelligence to some extent also helping the computers to become more intelligent by scrutinizing the non-verbal signals or sentiments of humans. An essential component of human-computer interaction is emotion recognition, which has attracted a lot of interest recently because of its potential uses in a variety of industries, including psychology, business, neuromarketing strategy, education, and entertainment. In this research, we propose a combination of Convolutional Neural Networks (CNNs) and XGBoost algorithms on Electroencephalogram (EEG) spectrogram images to propose an intriguing fusion-based model for identify four different classsed emotion, namely happy, sad, fear, and neutral. It has been researched that EEG signals hold important information about emotional states, and spectrogram images offer a good way to visualize this information. Before feeding the spectrogram images into the CNN-XGBoost model, Before transforming the EEG data to RGB pictures, we first use a Short-Time Fourier Transform (STFT). The XGBoost method is utilized for multiclass classification, while the CNN retrieves pertinent features from the spectrogram images. On our benchmark dataset called SEED-IV dataset which is publically accessible dataset for EEG-based emotion identification, our proposed approach was validated and it exhibited top-of-the-line precision and F1-score results. To do this, we extracted features from the signals using a range of feature extraction approaches, including the Short Time Fourier Transformation, Discrete Cosine Transformation, Power Spectral Density, Differential Entropy factors, and certain statistical traits. In order to demonstrate that the model we suggest is better in terms of accuracy and computing efficiency, we also conducted comparisons with a number of other well-known models. The performance analysis demonstrates that the suggested CNN-XGBoost fusion approach, which is based on spectrogram images, excels over conventional feature-based CNN, LSTM, and various pretrained models, including the VGG16 and VGG19 methods. Our stated CNN-XGBoost fusion-based framework using EEG spectrogram images delivers a promising method for precise and effective multiclass emotion identification, which has significant ramifications for facilitating the development of future systems for human-computer interaction.

Keywords: Emotion, EEG, Spectrogram, CNN, XGBoost, AI, Affective-computing, Human Computer Interaction.

Acknowledgement

The merciful Almighty Allah—the Most High, the Most Benevolent, and the Most Merciful—deserves all praise for having granted us the opportunity to study for a bachelor's degree and for completed our thesis without any major interruption here at BRAC University. It would not have been possible for us to travel the path to earning a bachelor's degree without the assistance of many people. We would like to take this opportunity to thank them for their encouragement and guidance.

We appreciate Md. Golam Rabiul Alam, PhD Sir, our supervisor, and Rafeed Rahman Sir, our co-supervisor, for their kind assistance and counsel in our endeavor. Without their guidance and invaluable time we could not have progressed with this project.

Finally, we should thank our parents because we might not have been successful without their constant support. We are now just a few steps away from graduating thanks to their kind prayers and support.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

AI Artificial Intelligence

BCI Brain-Computer Interaction

CNN Convolutional Neural Network

DCT Discrete Cosine Transform

DE Differential Entropy

DFA Discriminant Function Analysis

DL Deep Learning

EEG Electroencephalogram

FFT Fast Fourier Transform

$fMRI$ Functional Magnetic Resonance Imaging

HCI Human-Computer Interaction

LDA Linear Discriminant Analysis

LDS Linear Dynamic System

$LSTM$ Long Short-Term Memory

ML Machine Learning

$mRMR$ Minimum Redundancy Maximum Relevance

NN Neural Networks

PCA Principal Component Analysis

PSD Power Spectral Density

$RMS-Prop$ Root Mean Square Propagation

RNN Recurrent Neural Network

$SEED-IV$ SJTU Emotion EEG Sub-Dataset

SSL Semi-Supervised Learning

$STFT$ Short-Time Fourier Transform

SVM Support Vector Machine

VGG Visual Geometry Group

$XGBoost$ Extreme Gradient Boosting

Chapter 1

Introduction

Emotion recognition plays a significant role in human communication and interactions with both individuals and technology. While humans possess the ability to infer emotions through various cues, such as facial expressions, body language, and verbal communication, machines lack this innate understanding [1]. However, as technology continues to become an integral part of our daily lives, the field of emotion recognition, particularly through brain signals, has gained substantial attention. The goal of this field of study, known as Brain-Computer Interfacing (BCI), is to harness brain activities to communicate with and operate devices without using conventional neuromuscular channels.

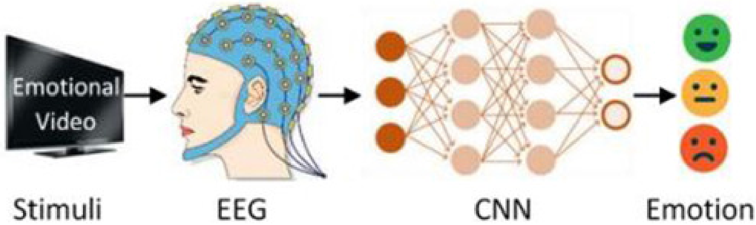


Figure 1.1: Emotion Classification From EEG

Over time, BCI systems have undergone continuous improvement, leveraging advanced techniques like feature extraction, band extraction, scaling, and classification models. The integration of wearable devices into everyday activities has further enhanced the feasibility of BCI systems. Electroencephalogram (EEG) signals have been widely used for emotion identification. However, challenges persist in inducing emotional states and extracting optimal traits for accurate classification [2]. Additionally, the reliance on expensive medical-grade equipment [2] limits the practicality of employing EEG signals for daily emotion analysis.

Although neuro network-based methods using visual and auditory inputs have been investigated, they often have accuracy issues since human emotions are inconsistently expressed via speech and facial expressions. Moreover, these techniques prove inadequate for individuals who have difficulty articulating their feelings or engaging in full interaction, particularly those with physical disabilities [3]. In this study, our objective was to enhance the detection rate of emotion recognition using a custom Convolutional Neural Network (CNN) model. We employed the SEED-IV [4] dataset, which contains a vast collection of EEG signals, and through our proposed methodology, we achieved substantial improvements over the original detection results.

Our approach involved the development of a custom CNN model tailored for emotion recognition. We trained the model utilising sophisticated methods, including extraction of features, band extraction, expansion, as well as classification models, by using the SEED-IV dataset [4]. Through extensive experimentation and refinement, we successfully improved the detection rate, yielding highly accurate results. The contributions of our research are significant in the field of emotion recognition. By utilizing the SEED-IV dataset and employing a custom CNN model, we have achieved remarkable enhancements in the detection rate, surpassing the results of previous methodologies. These advancements have the potential to impact various applications, including human-computer interaction, healthcare, and assistive technologies.

In the following sections, we will delve into the related works in emotion recognition, discuss our proposed methodology in detail, present the results and analysis of our experiments, and conclude with an overview of the contributions and future prospects of our research.

1.1 Problem Statement

In numerous disciplines, including psychology, business, neuromarketing strategy, education, and entertainment, emotion recognition is fundamental. But it's still difficult to effectively discern and categorize emotions from non-verbal cues and human expressions. Traditional methods frequently rely on feature-based techniques, which might not fully capture the complexity of emotions. Additionally, although research on utilizing EEG spectrogram images for multiclass emotion classification has been limited, the use of Electroencephalogram (EEG) signals for emotion detection has received attention.To address these issues, an improved fusion-based model is required to evaluate EEG spectrogram images, combining the advantages of Convolutional Neural Networks (CNNs) with XGBoost algorithms. The model can extract pertinent characteristics from the spectrogram images by incorporating CNNs, while XGBoost techniques enable multiclass classification. This fusion-based method presents a promising way to enhance the precision and efficacy of emotion classification. Furthermore, there is a research void in the thorough investigation and assessment of different feature extraction methods on EEG spectrogram images. Methods such as STFT, DCT, PSD, DE factors, plus statistical factors have the potential to enhance classification performance by removing essential information from the signals. For emotion recognition methods to advance in the context of human-computer interaction, several restrictions must be addressed. The outcomes of this study will aid in the creation of systems for multiclass emotion recognition that are more precise and effective. In turn, this will result in enhanced user experiences and interoperability between humans and machines across a range of industries, including business, education, entertainment, and neuroscience-based marketing.

1.2 Research Contributions

The purpose of this study is to develop a novel fusion-based approach for multi-class emotion classification using convolutional neural networks (CNNs) and XG-Boost algorithms. The main goal is to accurately detect four different emotion classes—happy, sad, fear, and neutral—by using electroencephalogram (EEG) spectrogram analysis. Due to its numerous uses in psychology, business, neuromarketing strategy, education, and entertainment, emotion recognition has attracted a lot of interest as an essential part of human-computer interaction. This research intends to improve the intelligence of computers by analyzing non-verbal cues and human emotions and utilizing the potential of AI-based emotional computing.In order to fulfill the objective of the study, it will first investigate how well spectrogram images can represent emotional states and how EEG signals can be used to encode these emotions. The benchmark dataset SEED-IV, a freely available dataset for EEG-based emotion recognition, will be used to verify the proposed method. Precision, F1-score, accuracy, and computational efficiency will be used to assess the CNN-XGBoost fusion-based model's performance. In order to prove the superiority of the suggested strategy, a comparison will also be made with other well-known models, such as feature-based CNN, LSTM, VGG16, and VGG19. This study intends to contribute to the development of accurate and powerful multiclass emotion recognition techniques utilizing EEG spectrogram pictures by addressing these research objectives. The preliminary results of this study have important ramifications for developing human-computer interaction systems in the future. The key contribution of this research are as follows-

- The following is a possible formulation of the study's goal: to provide a fusion-based strategy based on CNNs and XGBoost algorithms for precise multi-class emotion categorization utilizing EEG spectrogram analysis. Through the study of nonverbal signs and human emotions, the research intends to improve computer intelligence and advance the field of human-computer interaction systems by recognizing enhanced emotion recognition algorithms. With reference to precision, F1-score, accuracy, and computing efficiency, the suggested method will be tested using the SEED-IV dataset. To show how superior the proposed technique is, comparison analysis with other well-known models will be done. For future developments in this field, the preliminary findings of this study have significant ramifications.

- In the study, four emotions—happy, sad, fear, and neutral—are precisely identified using EEG spectrogram analysis. By using AI-based emotional computing to analyze nonverbal clues and human emotions, the research aims to improve computer intelligence. The SEED-IV dataset will be used in the study's verification process, and its performance will be evaluated based on the proposed model's precision, F1-score, accuracy, and computing efficiency. The superiority of the suggested approach will be shown by a comparison with various models (including feature-based CNN, LSTM, VGG16, and VGG19). In order to improve human-computer interaction systems, the project intends to build precise and effective multiclass emotion identification tools. The study's preliminary findings have a big impact on how this field will develop in the future.

- We used a fusion-based method that combined the XGBoost and CNN algorithms to enhance multi-class emotion classification. To represent and encode emotional states, EEG spectrogram analysis was used. The benchmark used to gauge the effectiveness of the suggested strategy was the SEED-IV dataset. The use of performance evaluation criteria include accuracy, precision, F1-score, and computing efficiency. To further emphasize the superiority of the fusion-based approach, a comparison was done with other well-known models. Overall, this study used a fusion-based methodology, benefited from EEG spectrogram analysis, made use of a benchmark dataset, applied performance assessment criteria, and carried out comparison analysis to support the development of precise and effective multiclass emotion detection methods.

- Based on EEG spectrogram analysis, the method used convolutional neural networks (CNNs) and XGBoost algorithms to precisely detect four different emotions (happy, sad, fear, and neutral). The created model made use of AI-based emotional computing's capability to examine nonverbal cues and human emotions. The goal of the project was to enhance emotion recognition to increase computer intelligence in human-computer interactions. Metrics for precision, F1-score, accuracy, and computing efficiency were used to assess the model's performance. To demonstrate its superiority, the created method was also contrasted with other well-known models as feature-based CNN, LSTM, VGG16, and VGG19. The goal of the study was to enhance precise and potent multiclass emotion identification algorithms across a range of industries, including business, neuromarketing, education, and entertainment.

- This study contributes in a number of important ways. First, it introduces a brand-new fusion-based method for precise multi-class emotion categorization using EEG spectrogram analysis, which combines CNNs and XGBoost algorithms. It has uses in business, neuromarketing, education, entertainment, and psychology as it increases the efficiency of emotion perception. By assessing non-verbal cues and human emotions, the study also illustrates the potential of AI-based emotional computing, hence increasing human-computer interaction systems. Thirdly, the research shows how well the generated model performed overall by comparing it to other recognized models and demonstrating its superiority. Overall, this research studies AI-based emotional computing, improves human-computer interaction systems, and develops emotion identification algorithms.

Chapter 2

Literature Review

The study of the brain and nervous system, as well as cognitive psychology, neurology, and neurophysiology, are all combined in the field of neuromarketing. Physiological response coming from external stimuli, and marketing, which investigates beneficial exchanges, can be used to explain how marketing influences customer and consumer behavior as well as purchasing and decision-making processes [5]. Moreover, neuromarketing uses different tools to detect customers' emotions or behaviors toward any product. From several studies it is pretty obvious that the use of multiple methods or tools, that is the multi-modal approach to recognise emotions is more effective than the unimodal method. Why our study is trying to focus on a multi-modal strategy in neuromarketing when there are numerous studies on this topic is one question that pops up. Simple answer is, multimodal has been utilized in a number of research, but there are several drawbacks that might be addressed with a modified, effective and more accurate model. As a result, the following section also gives a quick description of the improved multi-modal method's success in neuromarketing, especially when compared to earlier efforts.

2.1 Related Works

In this section, prior pertinent research will be critically reviewed in the fields of emotion detection and neuromarketing. Moreover, this section analyzes various methods employed to achieve the primary goal and provides a brief summary of multi-modal techniques used in neuromarketing in previous studies.

2.1.1 Fact-findings of commonly utilized technique of Emotion recognition

A brand-new technique for emotion recognition was put out by Zhao et al. [6] that employs RF signals to detect users' emotions, collects users' heart rate information, and creates an algorithm to recognize users' emotions.On the basis of brain-computer interface, Atkinson et al. [7] created an emotional trait recognition model based on EEG data [8].Additionally, MsEmoTTS, a numerous scales framework for emotional language synthesis that is inspired by the hierarchical structure of prosody, was suggested in a research paper. MsEmoTTS enables modeling of emotion at various levels and various approaches to emotional speech synthesis. Numerous experiments demonstrate that MsEmoTTS effectively creates emotional speech by anticipating the feeling from text entered or simply transferring the emotion from reference audio, and they also demonstrate that MsEmoTTS is able to be customized to create emotional speech as desired [9].

Here is a summarized table of some approaches and results for emotion recognition.

Table 2.1: Emotion recognition different approach and successes

Methods	Performances
Using CNN-RNN and C3D Hybrid Networks, video-based emotion recognition [10].	Achieved accuracy of 59.02%, which is the highest level and without adding any more video clips with labels for emotions in the training set [10].
HoloNet: Towards Accurate Wild Emotion Recognition [11].	57.84% on average recognition rate was attained [11]
Technique for Semi-Supervised Learning (SSL) [12].	is at least 5.0 percent more successful than traditional SSL techniques (absolute gain) and performs well in the high or low levels feelings of arousal categorization (UAR = 76.5%) [12].
Data-driven approach to investigate evidence of patterns of different emotion classes [13].	Achieved a UW accuracy of 65.60%, which is 1.90% better than the baseline [13].
EEG-based emotional models employing transfer learning methods without annotated selected data [14].	Positive emotions are more frequently recognized (85.01%) than other ways, while neutral (25.76%) and negative (10.24%) feelings are frequently mixed up [14].

2.1.2 Studies conducted using a multimodal approach to emotion recognition

A study was carried out on searching for a solution on mental health issues. It was suggested that users utilize an emotion tracking software to better understand and control how they feel. There is a virtual assistant in this program that, when end user are feeling down, will converse with them and offer upbeat remarks and suggestions for enhancing feelings based on previous interactions. Additionally, It displays the end user's sentiment indicators throughout each week in order that they may see whether they felt altered during the week. The program may also determine a user's emotions from their social media posts and display the positive and negative aspects of their online image. Ionic Framework and Python Flask were used to construct the application [15]. EmoCo, an interactive visual analytical system, was suggested as an alternative paradigm to examine emotion coherence in presentation videos across several behavioral modalities. Users may examine emotions at three distinct levels of detail using this system's five interconnected views (i.e., at the clip, phrase, and letter tiers) [15]. To facilitate visual study of movies, it combines cutting-edge designs with tried-and-true visualization approaches. To make it easier to explore multimodal emotions and their connections inside a movie, in this area of research, specifically suggest an augmented Sankey the clustering-based extension approach for observing the time evolution, and an illustration for analyzing sentiment harmony. In this study [16] they offer a multi modal framework like the above diagram. Our system's capacity to provide effective and insightful analysis is demonstrated by two usage scenarios based on interviews with two domain experts and TED Talk recordings [17]. A sensation recognition method constructed around constantly monitoring ECG and PPG data was proven in a different study [18]. These indicators are more widely available, easier to monitor, and more pleasant than other signals like blood pressure and EEG. The best collection of features in the period and amplitude domains were identified and retrieved to give an initial 48-dimensions input of information for an CNN framework featuring four layers of convolution and a speedy subject-dependent learning cooperation. The performance is on par with or better than comparable efforts, even with only two indicators being monitored. The aforementioned emotion detection system may be further incorporated at a computer-on-chip solution for actual time peripheral learning and categorization off-edge. [18].

2.1.3 Works on Neuromarketing and Emotion AI

There is a literature review on neuromarketing and AI, as well as the present application of AI in marketing studies using various technologies. Additionally, they have chosen to investigate the effectiveness of facial expression identification utilizing AI-based tools in extracting and analyzing consumers' emotions in response to marketing stimuli. They also mentioned the employment of vision tracking systems in upcoming projects [19]. The feature selection strategies used for neuromarketing priority detection approaches were described in a later piece of work. They evaluated how EEG indices affected the preferred state and spoke about several computational intelligence feature selection methods. Four methods—PCA, ReliefF, mRMR, and RFE—were employed to choose EEG features. All classifiers' performance was found to be enhanced by feature selection for EEG signals [20]. Additional research has described the exact methods employed in neuromarketing. Eye-tracking, EEG, and fMRI are some of the most popular methods used in academic and market research, respectively. Additionally, it is noted that neuromarketing is unlikely to ever completely replace conventional tactics, but that it does and will continue to play a crucial complementary function to the current qualitative and quantitative methodologies [5]. After evaluating several studies we can draw the inference that there are many studies regarding human emotion detection and neuromarketing. However, no study has been done using a multi-modal tool that consists of four parameters (such as,facial expression, speech data, EEG signal, and eye tracker), which has also been verified with the aid of a likert scale or valence arousal model. There are some observations of the application of multimodal approaches to recognize human behavior. We believe that our multi-modal strategy will be able to address many of the issues with neuromarketing that have been raised in earlier research.

2.1.4 Review on SEED-IV Dataset

Both internal physiological reactions and external acts are ways that emotions are expressed. Several signal modalities provide various facets of emotions, and complementing data from a variety of modalities may be combined—instead of using unimodal techniques—to create a system for more accurate emotion recognition. This paper introduces EmotionMeter, a multimodal tool for electroencephalogram and vision analysis of motion are used to determine feelings among individuals. Comfort and practicality have been taken into consideration when designing a six-electrode location above the ears that can be installed in a portable headpiece.When compared to a single modality (70.33% for EEG and 67.82% for eye movements), Modality fusion, which blends multimodal deep learning with EEG and eye movements, has been shown to greatly improve the accuracy of emotion categorization (85.11%) It has also looked at how eye movements and EEG work together to recognize emotions, as well as how consistently the recommended method works over time. The quantitative evaluation's findings proved that the suggested EmotionMeter tool works well [21], [4].

2.1.5 Fact-findings of different machine learning techniques of feeling identification

Numerous papers are conducted study on the human feeling identification among them this paper [22] on the subject of feeling identification from electroencephalogram signals, with an emphasis on different ML methods such as, an artificial neural network which is called deep learning and another method they used which is characterized by recalling and root memorization called shallow learning methods. The authors emphasize the value of EEG-based feeling identification in enhancing the field of HCI systems and its uses in cognition, affect detection, and sensation detection.The paper attempts to give a systematic assessment of the procedures and techniques used in recent studies in order to address the growing interest in this topic. Modern emotion identification systems reported in the literature are thoroughly examined by the writers. To help researchers create successful frameworks for emotion recognition systems, they provide an overview of the common procedures involved in emotion recognition, along with pertinent terminology, theories, and analyses. An artificial neural network based and memorization and recall based approach are the two basic categories used to group the systems [22]. The research compares these systems based on the techniques and classifiers utilized, the quantity of categorized emotions, accuracy levels, and the datasets used. The authors also give details regarding freely accessible datasets, analysis software, and tools, as well as a useful comparison table and performance graphs. The report finishes by reviewing current research trends in emotion identification from EEG signals and making suggestions for future research trajectories. The paper aims to support the development of efficient emotion recognition systems for real-life applications by compiling and analyzing existing studies and serving as a valuable resource for researchers interested in using deep learning or shallow learning techniques for EEG-based emotion recognition [22].

2.1.6 Affective state recognition using fusion model facts

Several research has been conducted on fusion based model for emotion recognition among them this study [23] deals with the problem of identifying emotional states in people using neuro biological signals. The authors provide a method for distinguishing between arousal (ease or excitement), valence (positive or negative mood), and dominance (helpless or in charge) [23]. This technique combines a convolutional neural network (CNN) with Extreme Gradient Boosting (XGBoost). DREAMER, a benchmark dataset that contains EEG signals as well as judgments of one's own performance, was employed in the study. The suggested procedure is employing Short-Time Fourier Transform (STFT) to turn EEG signals into spectrograms, which are then transformed into RGB pictures. The features are then extracted using a 2D CNN trained on these spectrogram images and fed into a dense layer of the neural network. The collected CNN features are then sent into the Extreme Gradient Boosting (XGBoost) classifier, which divides the signals into various emotion dimensions. The authors use SVM and XGBoost classifiers to analyse the accomplishment of the proposed technique with feature fusion-based algorithms in order to assess it. The signals are subjected to a number of feature extraction techniques, including FFT, DCT, and statistical features. To categorize various emotion levels, feature-level fusion is carried out, and the fused features are then used in SVM and XGBoost classifiers. According to the results of the performance analysis, the CNN-XGBoost fusion technique based on spectrogram images achieves better results than the feature fusion-based SVM and XGBoost approaches [23]. The suggested approach detects human emotions with high levels of accuracy, scoring 99.712% for arousal, 99.770% for valence, and 99.770% for dominance [23]. The research concludes by presenting a unique technique of human feeling identification from electroenchepalogram signals utilizing image spectrogram fusion using convolutional neural networks and extreme gradient boosting. The results manifest appreciable advancements over feature fusion-based methods employing a fusion based approach Support Vector Machine and Extreme Gradient Boosting classifiers. This recommended method appears to have a high likelihood of success in determining human emotional states, outperforming other experiments done on the same dataset. The authors propose that, in order to further improve performance, future studies should examine particular deep neural networks and other psychological cues [23].

Chapter 3

Description of Dataset

3.1 Description of Dataset

Zheng and Lu created the SEED dataset [24]. In [25], the SEED-IV dataset was originally suggested. From a variety of sources, 15 Chinese movie clips representing the emotions of happiness, neutrality, and sadness were selected as the stimuli for the research. The subjects were informed of the complete experiment's methods prior to the assessment. Participants were instructed to view that the 15 chosen footage during the experiments as well as to express their emotions. Four different emotions—happiness, sadness, neutrality, and fear—as well as eye movement signals were tracked in the dataset [21]. In all, 15 participants (7 males and 8 women) took part in the tests. Each subject had three sessions across a number of days, with every session containing 24 trials [21]. The participant watched one of the movie snippets for each trial. The participants were given 45 seconds to offer input following the viewing of a movie clip, followed by 15 seconds of respite. For the comparison research in this publication, We utilize the identical SEED dataset subset as in our recent chapter[26], [27], [28]. Electroencephalogram and vision fluctuation signals can be found in the SEED dataset [4]. Using a 62-channel electrode cap and an ESI NeuroScan technology that sampled at a rate of 1000 Hz, the neuro biological data were acquired[21]. Eye movement signals were collected using SMI eye tracking glasses. The trial's timetable is shown in the section below:

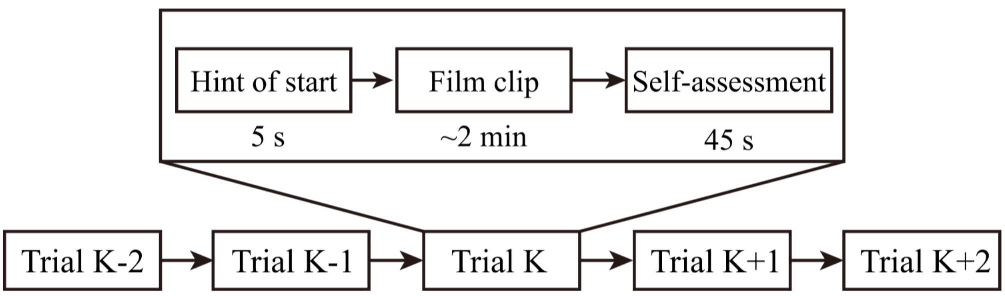


Figure 3.1: Protocol trial.

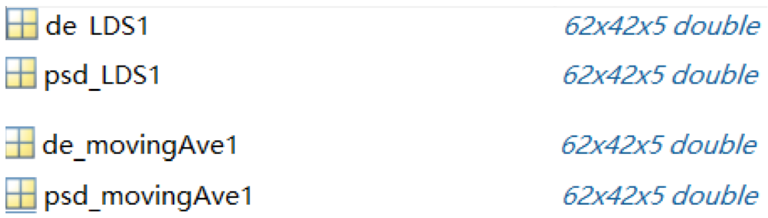


Figure 3.2: EEG feature folder structure.

Table 3.1: The electroencephalogram (EEG) data dimensionality for the SEED-IV datasets.

Hallmark	**Volume**
No. of entrant	15
Volume of Stimuli	72
Video footage	Video, audio
length of the footage	120 s
Entrant age	20 - 24
Recorded wave	EEG
Channels	62
Sampling rate	1000 Hz
Score of	Valance and arousal
Variable scoring scale	-5 to 5
Total data	$(15 \times 24 \times 3) = 1080$
Spectrogram images	$(15 \times 24 \times 3 \times 5) = 5400$
Train images	3780
Test images	1620

Chapter 4

Methodology

4.1 Methodology Workflow

4.1.1 Spectrogram Feature Based Methodology

Data Collection: The data for EEG signals is collected for emotion recognition purposes. Signal Processing: The collected EEG signals undergo signal processing techniques to prepare them for further analysis. Generate Spectrogram Images: Spectrogram images are generated from the preprocessed EEG signals. These spectrograms provide a visual representation of the signal's frequency content over time. Feature Extraction: Short-Time Fourier Transform (STFT) is applied to extract features from the spectrogram images. This process helps capture relevant information about the frequency components present in the signals. Convolutional Layers and Pooling: The extracted features are passed through convolutional layers, specifically Conv2D layers, to capture spatial information. Pooling layers are also used to reduce the dimensionality of the feature maps. Dropout and Batch Normalization: Dropout layers are introduced to prevent overfitting, while batch normalization helps normalize the input to each layer, enhancing the model's stability and performance. Dense and Flatten Layers: Dense layers are added to provide non-linear transformations and enable high-level feature learning. The flatten layer is used to reshape the output from the previous layers into a vector. Classification: The reshaped features are fed into classification models, including the Swin Transformer and XGBoost algorithms. These models perform the final classification of the spectrogram images into different emotion classes. Classification Results: The classification models generate results indicating the predicted emotions for the given spectrogram images.

In summary, the spectrogram data section involves data collection, signal processing, spectrogram image generation, feature extraction using STFT, convolutional layers, pooling layers, dropout layers, batch normalization, dense and flatten layers, classification using the Swin Transformer and XGBoost models, and obtaining the classification results.

4.1.2 Signal Feature Based Methodology

Data Collection: EEG signals are collected as the input data for emotion recognition. Signal Processing: The collected EEG signals undergo signal processing techniques to prepare them for further analysis.Band Pass Filter: The signals are filtered using a band pass filter to isolate specific frequency bands such as Delta, Theta, Alpha, Beta, and Gamma. Feature Extraction: Features are extracted from the filtered signals. Techniques such as Fast Fourier Transform (FFT), Discrete Cosine Transform (DCT), Differential Entropy (DE), and Power Spectral Density (PSD) are applied to capture relevant information about the frequency and power characteristics of the signals. Classification: The selected and reduced features are fed into classification models, including VGG16, VGG19, custom CNN, and LSTM networks. These models are responsible for classifying the input signals into different emotion categories. Classification Results: The classification models generate results indicating the predicted emotions for the given input signals.

In summary, the signal data section involves data collection, signal processing, band pass filtering, feature extraction using techniques such as FFT, DCT, DE, and PSD, feature selection, LDA for dimensionality reduction, classification using models like VGG16, VGG19, custom CNN, and LSTM, and obtaining the classification results.

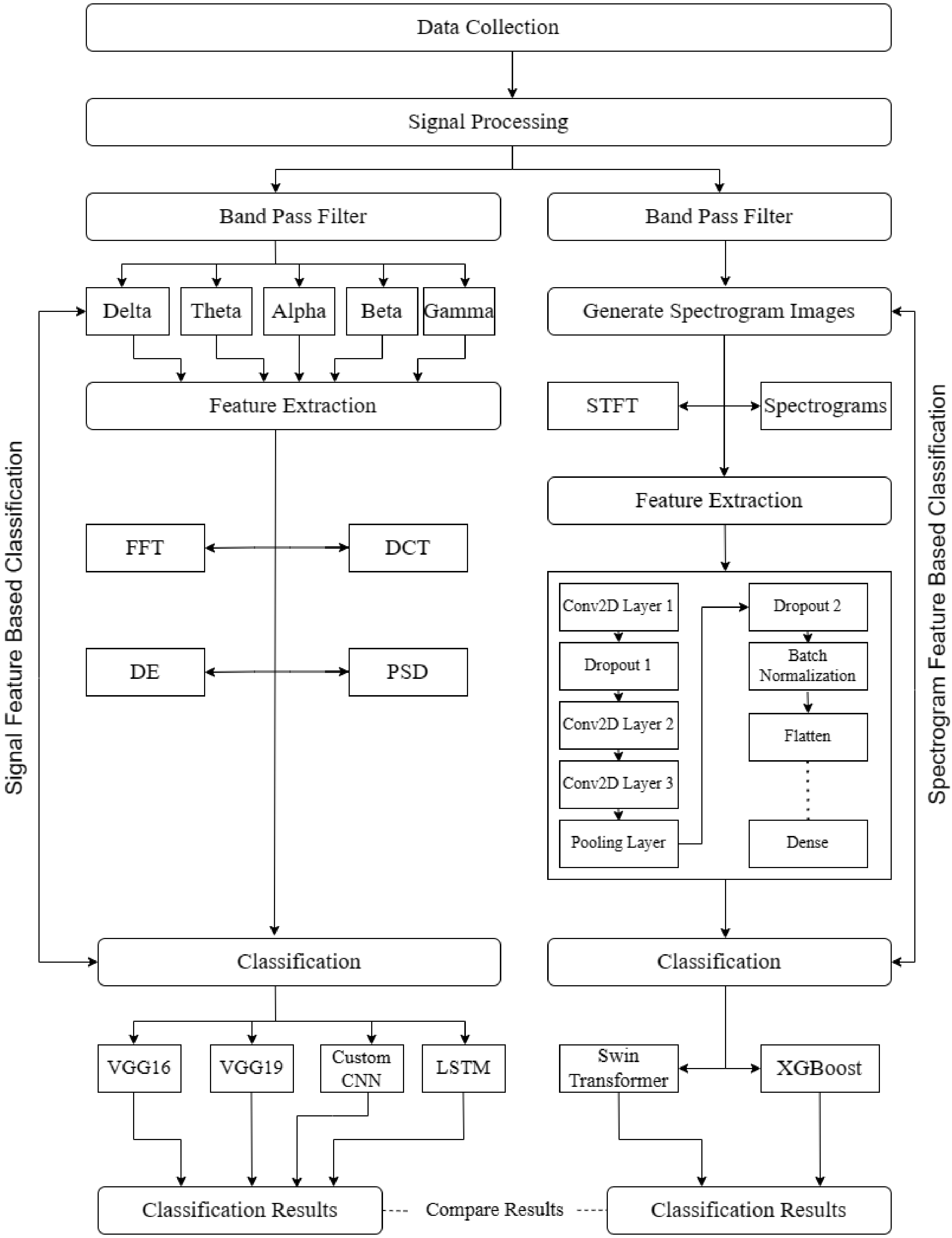


Figure 4.1: An overview of proposed CNN-XGBoost fusion approach.

4.2 Data Pre-Processing

We have pre processed our data in various ways below the brief description is given.

4.2.1 Signal Pre-Processing

Signals from electroencephalograms frequently contain a great deal of disturbance. Therefore, a high percentage of ocular distortions occur beneath 4 Hz, muscle contractions take place across 30 HZ, and electrical noise from the line take place between 50 and 60 Hz [2]. It is necessary to reduce or get rid of the noise for a more accurate analysis. In order to focus on a particular location, we must also pay attention to the frequency range that gives us the signals brought on by stimuli. A frequency spectrum from 4 to 30 Hz contains the data related to the emotion recognition test [2]. In order to distinguish the band of interest from the electrical noise in the waves, we used band pass filtering to gather sample values in the range of 1–50 Hz.

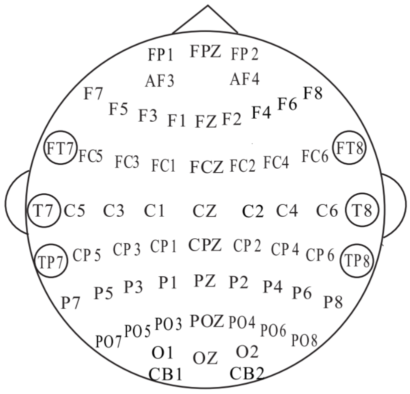
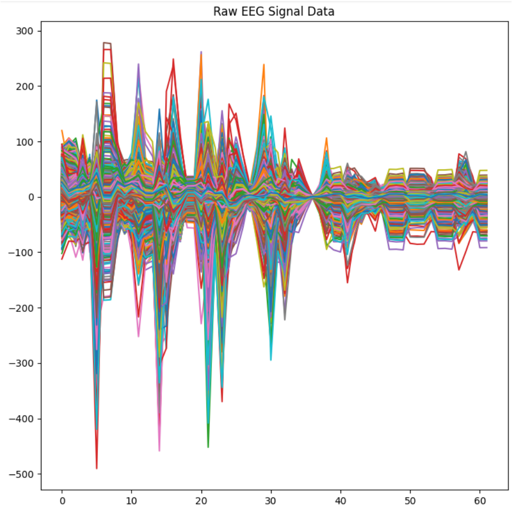


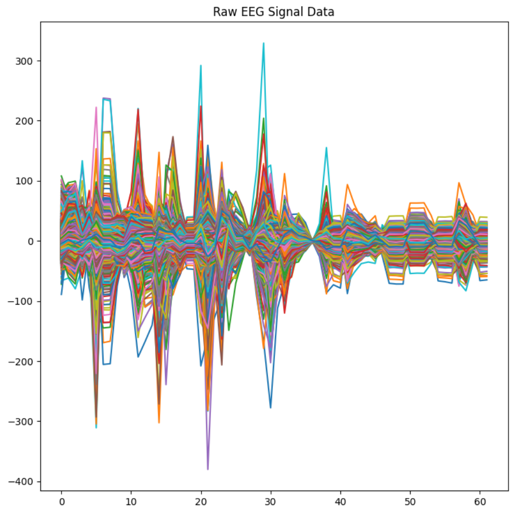
Figure 4.2: EEG electrode layout of 62 channels

4.2.2 Pre-process using the technique Band Pass Filter

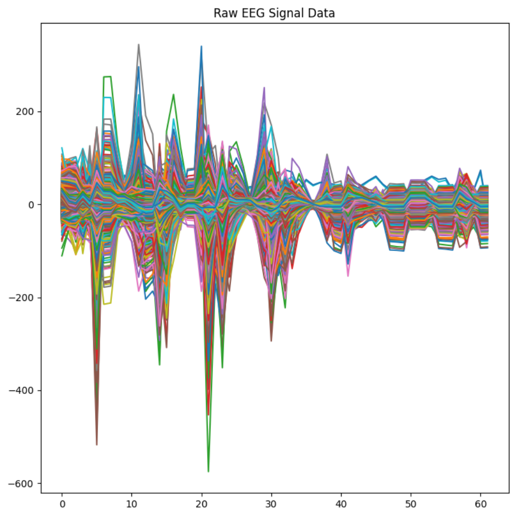
This technique are used to permit signals within some specific span of frequencies at the time rejecting those above or below the desired frequency. The band-pass filter combines the functions of a low pass and a high pass filter to filter out undesired frequencies. Such a filter's primary objective is to reduce the transmission range of the signal, enabling us to choose the information we want, whereas utilizing the required band of frequencies simultaneously cutting down on noise from bands we won't be using. Band-pass filtering was applied throughout the research approach in this study. Both the feature-based and the spectrogram-image-based fusion approaches employ this filtering method to remove noise in the 1-50 Hz range. This makes it easier in shutting out annoying noises. The five additional bands of delta, theta, alpha, beta, and gamma will now be used to separate the signals of interest. Because these are the bands that are most frequently employed for EEG signal processing, these bands were chosen. Band boundaries are relatively arbitrary to define. Our instance uses the following frequency ranges: delta between 1 and 4 Hz, theta between 4 and 8 Hz, alpha between 8 and 14 Hz, beta between 14 and 31 Hz, and gamma between 31 and 50 Hz. We also create spectrograms for the interval between the two values specified. In this case, spectrograms were retrieved from the data using STFT and then converted to Colourful pictures. Moreover, in this research we used two methods to make smooth the raw EEG data. The two smoothing methods are Linear Dynamic System (LDS) and Moving Average (movingAve). Filtering away noise and artifacts unrelated to EEG characteristics is done using these two separate methods.



(a) Trial 1



(b) Trial 2



(c) Trial 3

Figure 4.3: Raw EEG signal plotting before noise cancellation

4.2.3 Linear Dynamic System (LDS)

Systems are considered to have linear dynamics if their evaluation functions are linear. Dynamical systems don't typically have closed-form approaches, even though linearly dynamical systems can sometimes be exactly solved and include a variety of mathematical properties. The qualitative behavior of large dynamical systems may also be explained by linear systems by identifying the points of balance of the system and estimating it to be a linear system near each of those points. The Linear Dynamic System (LDS) uses the equation shown below.

$$X(t) = \sum_{k=1}^{n} (l_k \times X_0)^{r_k e_k t} \tag{4.1}$$

where, X_0 is the initial vector, r_k is the right eigenvector of matrix A and l_k is the liner combination of the right and left eigenvectors.

4.2.4 Moving Average (movingAve)

The moving average is an estimation tool for forecasting trends over time. In order to average a collection of data from a specific range, the procedure involves moving the range. The simple moving average (SMA) is a form of moving average (MA) used in prediction of time series. It is established by mathematically averaging data gathered over a set period of timeframe. It advances the window over a certain period of time. The Moving Average (movingAve) formula is as follows.

$$MA = (a_1 + a_2 + ... + a_n)/n \tag{4.2}$$

where, A is the average in period n and n is the number of time periods

4.2.5 Data Labeling

One person in our dataset participated in the experiment 24 times in one session, while the entire experiment was carried out over the course of three sessions [21]. In 24 trials the expatriates play four different emotional clips such as, happiness, sadness, fear, and neutrality randomly. every emotional video clip was played 6 times randomly and we label our dataset on this basis.The demonstration of the labeling is:

Label:
The labels of the three sessions for the same participants are as follows,

session1_label = [1,2,3,0,2,0,0,1,0,1,2,1,1,1,2,3,2,2,3,3,0,3,0,3];
session2_label = [2,1,3,0,0,2,0,2,3,3,2,3,2,0,1,1,2,1,0,3,0,1,3,1];
session3_label = [1,2,2,1,3,3,3,1,1,2,1,0,2,3,3,0,2,3,0,0,2,0,1,0];

The labels with the numbers 0, 1, 2, and 3 stand for the truth, neutrality, sadness, fear, and happiness, respectively [21], [4].

4.2.6 One Hot Encoding

One-hot encoding converts categorical information into a format that machine learning algorithms may accept as input. This critical pre-processing stage is required for many machine learning initiatives. The goal of one-hot encoding is to convert material from a categorical form to a numerical representation. Following the ordinal encoding of the category variables, each integer value is represented as a binary vector with all other values of the vector being zero except from the integers' position, which is labeled with a 1. In other words, this is a method for turning a category feature with string labeling into K nominal attributes so that the value of the remaining (K-1) features is 0 and the outcome of one of the K (one-of-K) characteristics is 1. It is often referred to as ”dummy encoding” since the features generated as a result of these methodologies are dummy features that do not represent any real-world characteristics. They were instead created to employ fictitious numerical qualities to encode the different category feature values. Take a look at a dataset that contains information on several vegetable kinds. Each vegetable type would have its own column in one-hot encoding, with a value of 1 denoting that the rows contain information about that product and a value of 0 denoting that they do not.

One-hot encoding might be accomplished using the Pandas module for Python. Using the ”get dummies” function provided by the Pandas package, data may be one-hot encoded. Additional approaches are available, such as one-hot encoding using the OneHotEncoder class from the sklearn.preprocessing package. Data that has been one-hot encoded is also known as dummy data. Dummy data makes machine learning models easier to employ since it does not need special processing of categorical variables.

The categorical characteristics must be translated or turned into numerical features using the one-hot encoding method in order for machine learning libraries to use the values to train a model. Although many machine learning frameworks already accomplish this internally (as fake features), it is still advisable to explicitly translate this category data into numerical features.

4.2.7 Converting dataset from .mat format to .csv format

In our work we generate the spectrogram from the data for our convenience. We transform the dataset into Comma-separated values (CSV) file format. For that we used Python's scientific computation library SciPy, Pandas as a dataframe and NumPy. First of all, we load our .mat file and convert that into the .csv format with the help of Pandas dataframe.

4.3 Feature Extraction

In this section, what kind of feature extracted from the data will be critically describe with the methods. It provides a brief summary of the feature extracted from the data.

4.3.1 Fast Fourier Transformation (FFT)

One of the most beneficial techniques for processing diverse signals is the fast Fourier transform [29],[30],[31], [32],[33]. We used the fast fourier transform method to compute a set of discrete fourier transforms. The FFT roots are evaluated according to their ability to function in the frequency domain, whether in the temporal or spatial dimensions. The FFT may also be used to get the O(NlogN) answer, where

N represents the vector's magnitude. A signal in the N time domain is split into N individual signals in a single step. In the second step, we estimate the periodicity band of the N in the time-domain signals. The N spectral was finally synthesized to produce a single- frequency constant to expedite the convolutional neural network's learning process. The FFT equations are shown in the following

$$H(p) = \sum_{t=0}^{N-1} r(t) W_N^{pn} \tag{4.3}$$

$$r(t) = \frac{1}{N} \sum_{p=0}^{N-1} H(p) W_N^{-pn} \tag{4.4}$$

where, H_p represents the Fourier co efficients of $r(t)$.

4.3.2 Mobility

This variable either indicates the median frequency or the spectrum's fractional natural dispersion. Its formula is the value of the square root for the dispersion of the y(t) signal's first derivative divided by y(t). This is the action's notation:

$$\sqrt{\frac{var(y^{‘}(t))}{var(y(t))}} \tag{4.5}$$

4.3.3 Complexity

The frequency shift is reflected by the parameter. Te compares how closely the signal resembles a pure sinusoidal wave, with a value of 1 if the signal is more similar. Below is a marker for activity.

$$\frac{mobility(y^{‘}(t))}{mobility(y(t))} \tag{4.6}$$

4.3.4 Statistical Features

Statistical is the branch of mathematics that deals with the processing of data for scientific or practical purposes. We focus on the mathematical outputs of our work with information-based data with statistical features. By studying statistics, we may have a better understanding of the methods that can be used in conjunction with this data to provide more reliable and well-structured results in the field of data science. Statistical methods have been used in several research of emotion analysis [34], [35], [3].

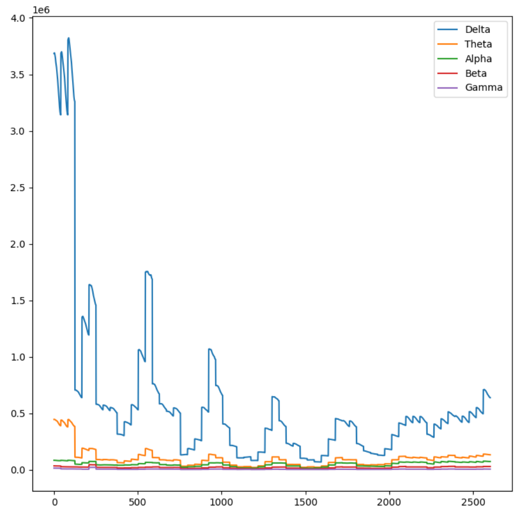
4.3.5 Power Spectral Density (PSD)

The average periodogram is evaluated using the Welch technique, a modified segmentation system, which is utilized in articles. A time series is subjected to the Welch technique. Its main goal is to lessen the fluctuation in spectral density readings.PSD identifies the substantial frequency range differences and may be very helpful for future study. The Welch approach of the PSD is commonly defined using the next power spectrum equations:

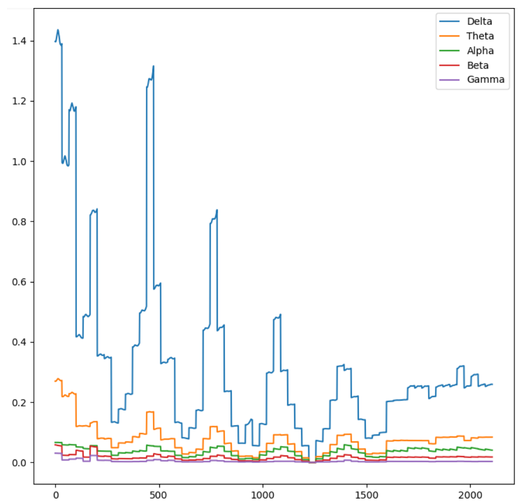
$$P(f) = \frac{1}{MU}\Big|\sum_{n=0}^{M-1} x_i(n)w(n)e^{-j2\pi f}\Big|^2 \tag{4.7}$$

$$P_{welch}(f) = \frac{1}{L}\sum_{i=0}^{L-1} P(f) \tag{4.8}$$

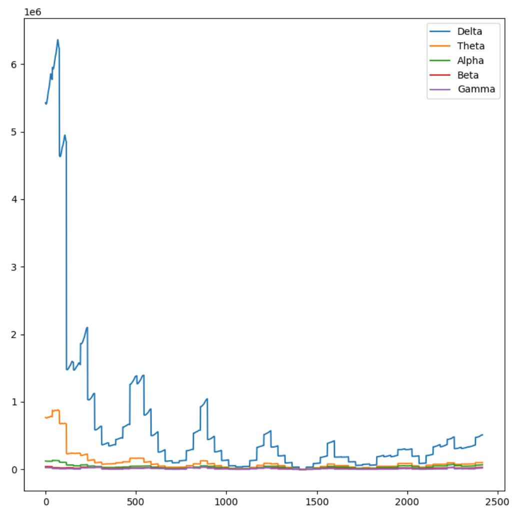
Here, the definition of the density equation comes first. Welch Power Spectrum then suggests that the average time is reported for each interval. To determine the PSD of the signal, we used the Welch technique. Each band's mean power has been determined from that.



(a) Trial 1



(b) Trial 2



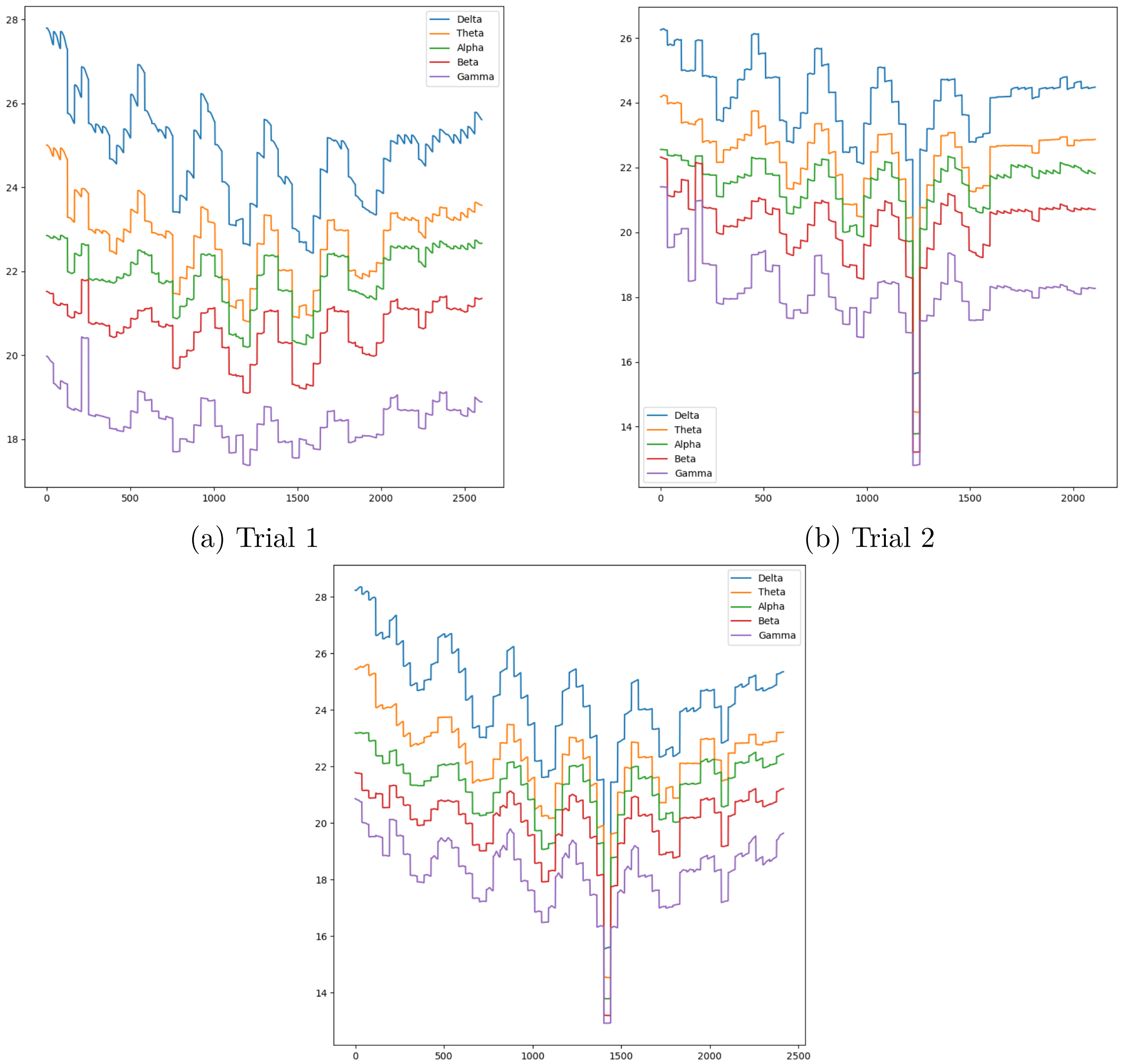
(c) Trial 3

Figure 4.4: EEG signals data after applying PSD

4.3.6 Differential Entropy (DE)

After downsampled the EEG data at 200 Hz sampling rate we extract differential entropy (DE) features within each segment at 5 frequency bands: 1) delta: 1-4 Hz; 2) theta: 4-8 Hz; 3) alpha: 8-14 Hz; 4) beta: 14-31 Hz; and gamma: 31-50 Hz [21]. The calculation of DE of a random variable x is [36], [37], [38]

$$DE = -\int_{-\infty}^{\infty} P(x) ln(P(x)) dx \tag{4.9}$$



(a) Trial 1 (b) Trial 2

(c) Trial 3

Figure 4.5: EEG signals data after applying DE

4.3.7 Discrete Cosine Transformation (DCT)

With this technique, each cosine value is represented by a set number of points over a range of frequencies; it is used in research [39], [40],[41], [42]. The Discrete Cosine Transformation (DCT) is often used to the coefficients of a recurring, proportionately stretched pattern in the Fourier Sequence. DCT is the most popular TDAC in the field of signal processing.. The signal's imaginary part is zero in both the temporal and spectrum spaces. Instead of being asymmetrical as in the mind's eye, the real part of the spectrum is more like a circle. The following equation may be used to convert between standard frequencies and mel frequencies.

$$X_p = \sum_{n=0}^{N-1} x_n cos\Big[\frac{\pi}{N}(n + \frac{1}{2})P\Big] \tag{4.10}$$

Where, N is the list of real numbers and X_p is the set of N data values.

4.4 Spectrogram Feature Extraction

Following data pre-processing, we employed a variety of feature extraction techniques for our CNN-XGBoost fusion-based strategy, which we will cover below, as well as our feature-based approach.

4.4.1 Short-Time Fourier Transform (STFT)

Although the time series is transformed using a Short-Time Fourier Transform (STFT) into a time-frequency diagram, a Convolutional Neural Network (CNN) is usually utilized to analyze pictures. It classifies pictures with fully linked layers after utilising multilevel pooling and convolution to extract relevant information from input photos. Using the filtered signal, which has a frequency range of 1 to 50 Hz, we computed the STFT and converted it into RGB pictures. In addition, we take the sampling frequency rate of EEG signal of 1000 Hz and the window size of 150 per segment. Below figures displays some of the produced photos. Time series EEG signals are often transformed using the wavelets techniques and FT, both of which we have included into our method. However, to maintain the accuracy of the source information, electroencephalogram transformation have to be done at the domain of period and frequency. Consequently, STFT is the best approach to preserving all possible anesthetic characteristics of the EEG data. The signal's spectrograms were retrieved using STFT, and the following equation is provided:

$$Z_n^{e^{(j\hat{\omega})}} = e^{-j\hat{\omega}n}[W(n)e^{j\hat{\omega}n} \times x(n)] \tag{4.11}$$

where, $e^{-j\hat{\omega}n}$ is the signal modulating the output of a complicated bandpass filter. Using the aforementioned formula, we determined the STFT of the filtered signals..

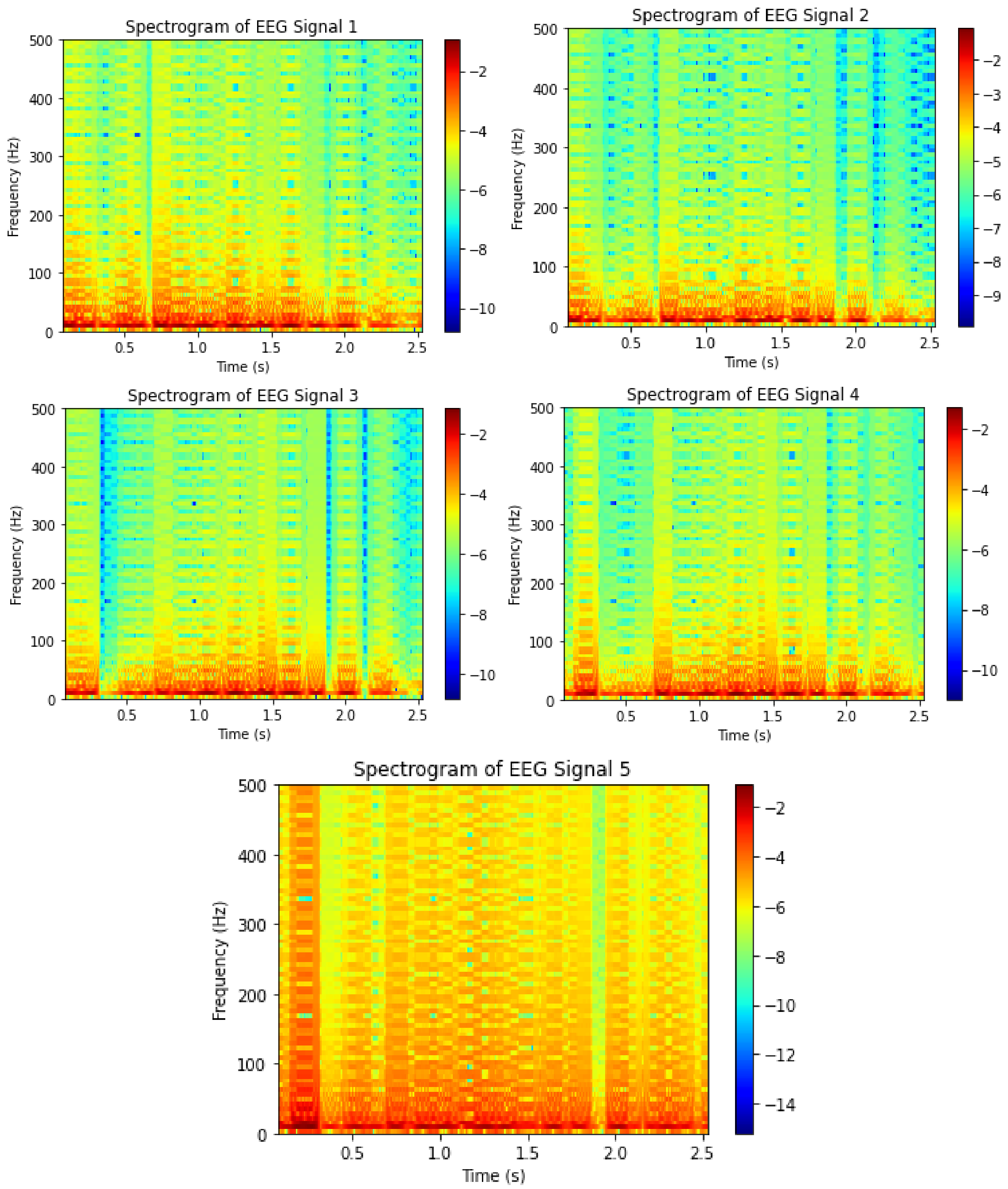


Figure 4.6: Spectrogram images generated from EEG signals

4.5 Signal Feature Based Classification Models

To achieve the highest accuracy we have implemented variety of machine learning algorithms. In this section, it will provides the brief summary of the algorithms.

4.5.1 VGG16

Convolutional Neural Network (CNN) model VGG16, also known as the Visual Geometry Group model, paved the way for several significant developments when it comes to visual computing. Karen Simonyan and Andrew Zisserman [43] first proposed the model in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2013. A pooling layer that reduces the image's height and breadth is followed by some of the 16 convolutional layers that make up VGG16.

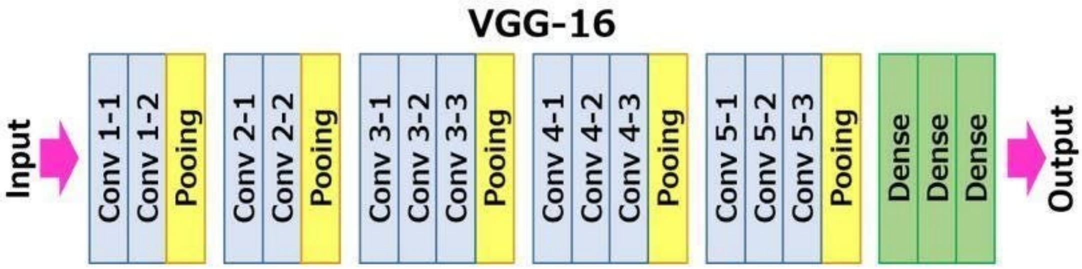


Figure 4.7: Basic Structure of VGG16

An adaptation of the VGG16 model architecture was used to create one of our emotion detection models.We have established the picture dimensions 224×224×3 in detail using the parameters derived from the already trained ImageNet network. In addition, a new Dense layer with softmax activation has been added in alternative of the original architecture's final Dense layer.

4.5.2 VGG19

VGG-19 has become the name of a convolutional neural network with 19 layers. A previously trained version of a network that was successfully learned upon over one million images is included in the ImageNet database [44]. A number of animals, a keyboard, a mouse, and a pencil are just a few of the 1000 distinct categories of products that the previously trained system can classify images into. As a result, the network now includes vast feature representations for a range of images. It will accept photos with a 224 by 224 resolution.
An adaptation of the VGG19 model architecture was used to create one of our emotion detection models. We are using the weights of the pre-trained ImageNet model. Provides input of fixed size of 224×224×3 RGB image. The only pretreatment done was to determine the average RGB value for every single pixel across the entire experimental collection. Used kernels of (3×3) size, This allowed them to completely hide the idea of the photo. Max pooling was carried out with stride 2×2 by 2 pixel frame. Following this, a rectified linear unit (ReLu) was employed to introduce nonlinearity into the model, improving categorization sensitivity and the speed of computation. This model performed quite superior to prior models that used tanh or sigmoid functions. Lastly, softmax activation has been added in alternative of the original architecture's final Dense layer.

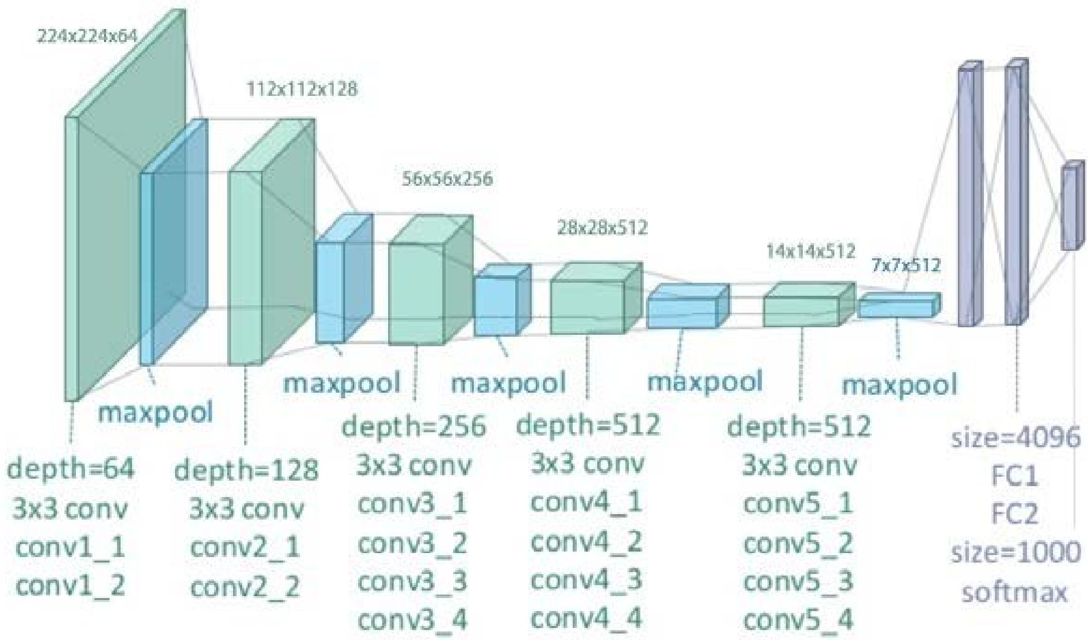


Figure 4.8: Basic Structure of VGG19

4.5.3 Neural Networks

Deep learning (DL), a branch of machine learning (ML), is built on neural networks, and its methods are modeled after the structures of the human brain. Neural networks essentially function by learning from data, identifying patterns, and then forecasting the results of a fresh collection of data.

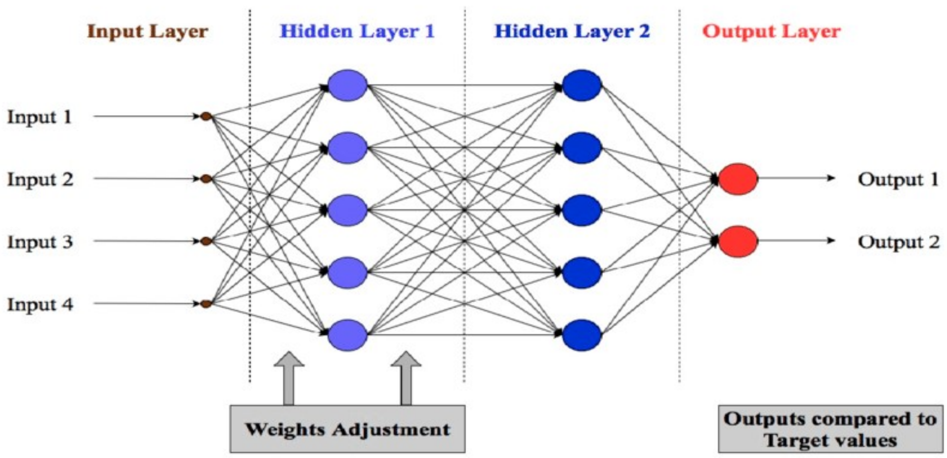


Figure 4.9: Basic Structure of Neural Network

The basic components of a neural network are a layer of input, a layer of output, and a number of layers which are hidden. After that input layer gets the inputs and the layer of output predicts the outcome, the hidden layers do the necessary computations to produce the output. Each layer contains neurons, it's called the network's main processing units. A weight, which is a numerical number, is assigned to each of the channels that link these neurons from one layer to another. After multiplying the data inputs by the respective weights, the total is provided as a parameter of input into the layer that is hidden. Every single one of those neurons is now once more coupled to the bias, a numerical value in that layer. Now, this sum is subjected to the activation function, a threshold function. The activation function determines the activation state of a particular neuron. The information is transmitted by the neurons in this manner all the way to the output layer. Forward propagation is the term for this transmission technique. The projected value in the output layer is ultimately chosen as the neuron with the highest value. There will unavoidably be errors, though, because they are only possibilities. Here is where the training phase comes into play. The errors and the actual outputs are compared, and the results are then communicated back to the network. In light of this, the weights and bias are adjusted.

4.5.4 Customized Convolutional Neural Network (CNN)

A computerized network of neurons that has been specifically trained to recognize or detect patterns is called a convolutional neural network, or CNN for short. CNNs are very valuable tools for doing image analysis as a consequence of this pattern discovery. The feature that distinguishes CNNs from other varieties of neural networks is the usage of convolutional layers, which are hidden layers in CNNs' terms. Convolutional neural networks (CNNs) are composed mostly of convolutional layers, while they may also typically incorporate extra layers that aren't related to the convolutional process. A convolutional layer is one that accepts input, alters it in some way, and then delivers the altered data to the layer below it. The phrases input channels and output channels are used to describe, respectively, the inputs and outputs of convolutional layers.

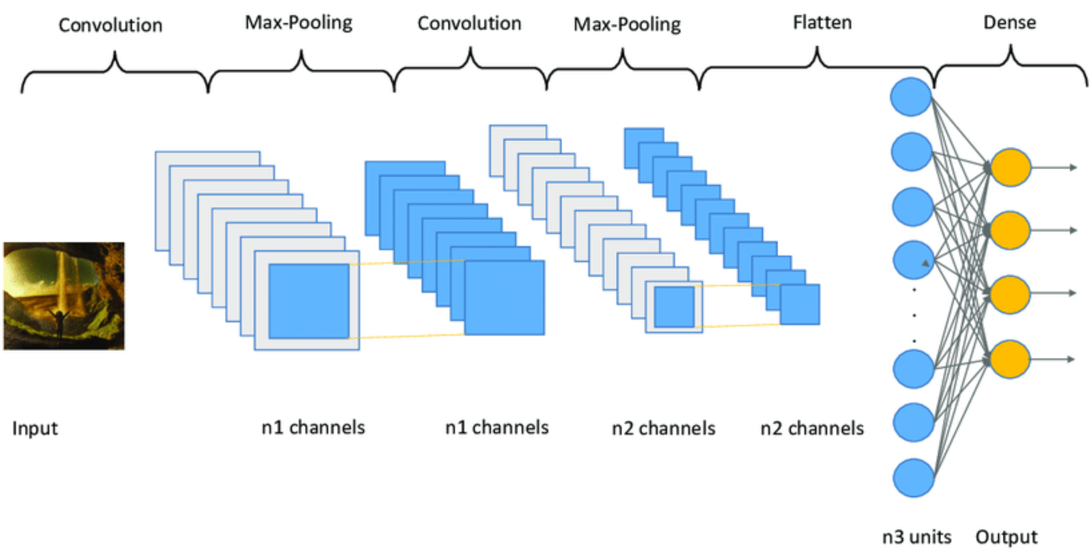


Figure 4.10: Basic Structure of Convolutional Neural Network

The change that occurs when dealing with a convolutional layer is known as a convolution operation. In any case, deep learning professionals refer to it in this manner. Cross-correlations, as they are known in mathematics, are carried out by convolutional layers. Convolution is a sort of operation that this is. To identify patterns in photos, one method is to employ convolutional neural networks. Every convolutional layer's necessary amount of filtration must be provided. The patterns are correctly found using these filters.

4.5.5 Long Short-Term Memory (LSTM)

The LSTM (long short-term memory) network is a typical RNN (recurrent neural network) type for learning consecutive data prediction tasks. Similar to other neural networks, LSTMs have a number of layers that assist learning and pattern recognition to enhance performance. The fundamental purpose of an LSTM may be understood as maintaining the data you require and rejecting the data that you require but will not be helpful for making future predictions. ANetworks based on LSTM come in a variety of forms, however they may typically be categorized into three categories. LSTM forward pass, LSTM reverse pass, Bi-LSTM. In contrast to bidirectional LSTM, which processes information on both sides to maintain it, the forward pass and backward pass LSTM are unidirectional, processing information in only one direction, either on the front side or on the rear. Each of the LSTM types previously mentioned functions in accordance with a fundamental framework. A basic LSTM model architecture may be broken down into numerous pieces.

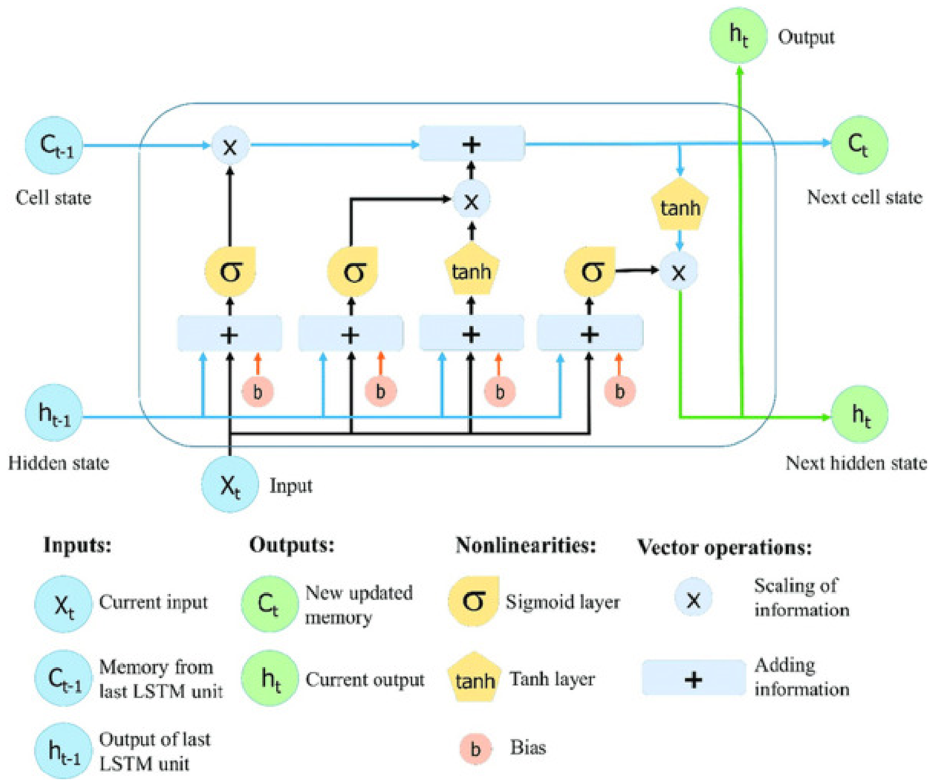


Figure 4.11: Basic Structure of Long Short-Term Memory

One of LSTMs' most notable characteristics is its capacity to remember and recognize incoming information and to reject that which is not required by the system to acquire facts and generate projections. This gate results in this LSTM attribute. It assists in deciding whether information can pass across the layers of the network.

Data through the previous layer as well as data from the presentation tier are the two different sorts of data entry that are anticipated through a system of network.

4.6 Spectrogram Feature Based Classification Models

4.6.1 Swin Transformer

One variety of Vision Transformer is the Swin Transformer. Hierarchical feature maps are produced by integrating image patches in deeper levels. Since attention to oneself estimations being limited to each local window, it has linear computing complexity for input image size. Thus, it may act as a general-purpose foundation for applications like dense recognition and picture categorization. As a result of computing self-attention globally, earlier vision Transformers construct mappings of features at one poor quality with a polynomial computationally demanding initial picture dimension.

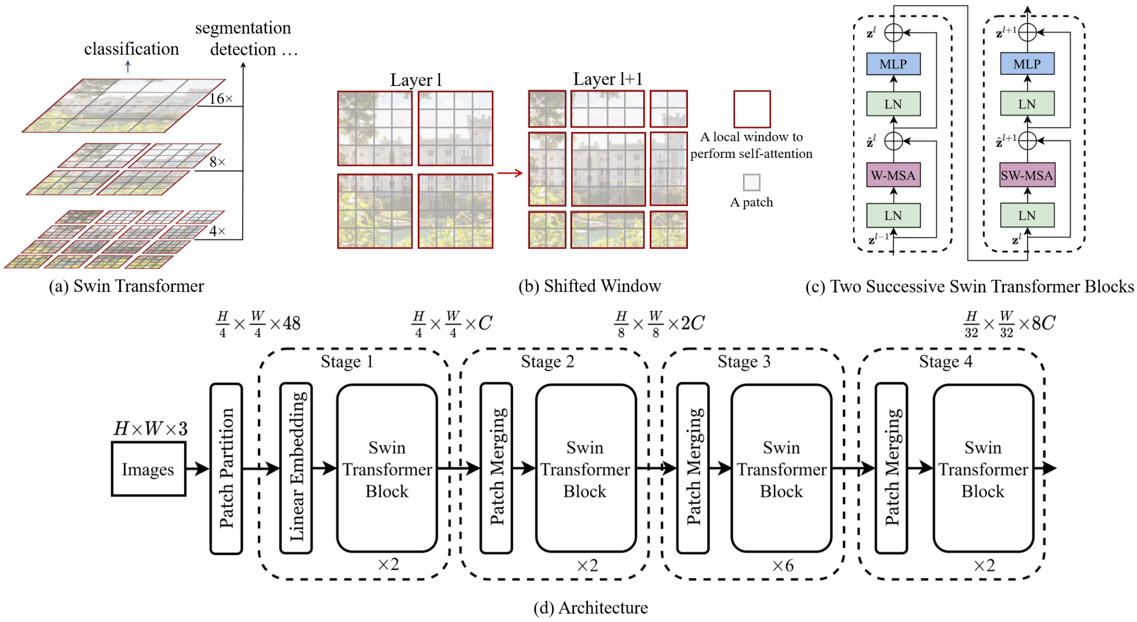


Figure 4.12: Basic Structure of Swin Transformer

Patches are still employed, similar to the ViT paradigm. Unlike ViT, where all patches are the same size (16×16px), Swin uses a variety of patch sizes in its first Transformer layer. The model merges these layers into bigger ones in the deeper levels. The image is broken up into patches that are each 4 pixels by 4 pixels. Each patch has a three-channel, vibrant image. Consequently, a patch has a 48-dimensionality total feature. Therefore, 4×4×3 equals 48. Then, it is sequentially transferred into the C-dimensional space, which is the one you choose. The output is then merged by the merging layer. Concatenating the vectors of numerous 2×2 neighbouring patch groups. The focusing window alters every time in respect to the layer preceding it. For instance, these parts are relocated in the subsequent layer (as in strided convolutional) if the initial layer's attention was solely on the area surrounding these areas. Patches that previously displayed in distinct frames and remained unwilling to interact are capable of doing that on the second layer. These created patches are combined in the layer of merging. This phase is repeated based on how many levels are chosen.

4.6.2 Extreme Gradient Boosting (XGBoost)

A machine learning approach called Extreme Gradient Boosting (XGBoost) combines the predictions of multiple weaker models in order to properly predict an objective variable. This serves as a standard, quick, high-performance data extraction method. The Extreme Gradient Boosting model computes ten times faster than the Random Forest model. The mixture tree approach was used to create the XGBoost model, which requires adding a new tree at every stage to go along with those that previously constructed. The accuracy often increases as more trees are created. After applying custom CNN to our suggested model, we applied XGBoost. From the trained layer of CNN, various characteristics were retrieved. Then, we classified all of the dimensions of emotion using Extreme Gradient Boosting on the basis of the returned features. Extreme Gradient Boosting employs the following formulae.

$$f(m) \approx f(k) + f^{'}(k)(m-a) + \frac{1}{2}f^{n}(k)(m-k)^2 \tag{4.12}$$

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n} \left[l(q_i, q^{(t-1)}) + r_i f_t(m_i) + \frac{1}{2} s_i f_t^2(m_i) \right] + \Omega(f_t) + C \tag{4.13}$$

where C represents as Constant, r_i and s_i are represents as,

$$r_i = \partial \hat{z}_i^{(b-1)} \cdot \int (z_i, \hat{z}_i^{(b-1)}) \tag{4.14}$$

$$s_i = \partial \hat{z}_i^{(b-1)} \cdot \int (z_i, \hat{z}_i^{(b-1)}) \tag{4.15}$$

The exact goal at step b changes once all of the constants have been eliminated.,

$$\sum_{i=1}^{n} \left[r_i f_t(m_i) + \frac{1}{2} s_i f_t^2(m_i) \right] + \Omega(f_t) \tag{4.16}$$

where,

$$\Omega(f) = \gamma P + \frac{1}{2}\lambda \sum_{j=1}^{P} Z_i^2 \tag{4.17}$$

4.6.3 A Customized Deep Convolutional Neural Network for Training

A particular kind of deep neural network called a convolutional neural network (CNN) is used in deep learning to interpret visual images. The CNN-XGBoost fusion method that we propose uses a general, two-dimensional Convolutional Neural Network model. Before employing this CNN architecture, Through restricting the wavelength band holding significant signaling, we were able to produce spectrum pictures, which is between 1 and 50 Hz. Next, we compute STFT of the electroencephalogram data, transmute them to pictures of spectrograms, and then retrieve attributes using a 2D convolutional NN. Then training the algorithm with 2D convolutional layers using the resulting spectrogram pictures. A dense layer is then used to retrieve the learnt features from the learning layer. The Convolutional Neural Network (CNN) described below which we suggest to train the proposed model.

4.6.4 Conv2D Layer

The most essential properties, such as edges which extend continuously and in a diagonal fashion are often extracted from initial layer. The following layer receives this knowledge and is in charge of identifying more intricate properties like edges and mixed borders. In addition, we explore the network further, it can detect increasingly complex items, including objects, faces, and other things. The layer used for classification generates a number of assurance scores (those between the which vary from 0 to 1) on the final convolution layer that indicate the probability that the picture corresponds to an identifiable category. We identified the classes in our proposed technique using three Conv2D layers.

4.6.5 Pooling Layer

Reducing the size of the space occupied by the convolved features is the responsibility of the pooling layer. By making the data smaller, the amount of processing power required is reduced. The two categories of pooling are average pooling and maximum pooling. Max pooling was employed because it produces better results than average pooling. It calculated highest possible score of a picture elemnet from a portion of the picture that the kernel covered using another layer called max pooling. It removes all spurious events, reduces dimensionality, and de-noises the signal. The following formula may be used to express any pooling function in general.

$$q_j^{(l+1)} = Pool(q_1^{(l)}, ...q_i^{(l)}, ...q_n^{(l)}), q_i \epsilon R_j^{(l)} \tag{4.18}$$

where, $R_j^{(l)}$ is the layer l's jth pooled area, and Pool() is a pooling function over that region.

4.6.6 Dropout Layer

In order to reduce overfitting, we added two dropout layers. When the reduction rate and dropout rate decline, the accuracy will continue to get better. Some of the maximum pooling's results are arbitrarily chosen and totally disregarded. They are not carried over to the next tier.

4.6.7 Flatten Layer

A flatten operation must always be carried out following a sequence of 2D convolutions. Data is flattened into a one-dimensional array so that it may be processed further. We flatten the output of the layers of convolution to produce a single long feature vector. It additionally has a connection to the general classification structure

4.6.8 Dense Layer

Dense offers the neural network a layer that is fully connected. Every the neurons with outputs from the layer above are sent to each neuron in the preceding layer, the layer above receiving a single outcome from every single neuron. With this CNN architecture, the approach we propose uses a variety of kernels in order to obtain key characteristics from the convolution layer. producing various feature maps. The CNN model has a completely linked layer at its conclusion. Then the last output of the fully linked layer creates the expected class labels for emotions. To extract this many features, we added a dense layer of 250 units after the training layer in accordance with our suggested methodology.

4.7 Hyper-Parameters and Auxiliary Functions

The model hyper-parameter and auxiliary functions are as follows

4.7.1 Kernel Size

The word ”step size” refers to the extent of the filter's sliding, whereas ”kernel size” refers to the size of the filter's orbit around the feature map. Maintaining a kernel size of 3×3 or 5×5 is standard in this situation. Without information of the surrounding pixels, the feature recovery would be local and fine-grained. The convolution layers' kernel size was set at 3.

4.7.2 Filter Size

Small matrices known as filters include weights or other characteristics that are learnt during training. Edges, textures, and forms are just a few examples of the patterns or elements in the picture that these filters are intended to find. The size of each filter is often smaller than the input picture. The filter size in the convolution layers was set to 64 and 32.

4.7.3 Padding

Padding is the number of dots that are included on a picture before the CNN kernel processes it, is advantageous for convolutional neural networks (CNNs). For instance, if a CNN's padding has been configured to 0, any extra pixels will be given the value 0. We didn't change the padding for our model kept it the same

4.7.4 Optimizer

Optimizers can be helpful when trying to get a model to run reasonably efficiently. They have traditionally been applied to improve the model's efficiency while lowering the inaccuracy caused by loss functions. These enable adjusting the weights and learning rate of a neural network to reduce losses and improve effectiveness. A model can benefit from a variety of optimizers to improve performance. Adaptive Moment Estimation (Adam), Root Mean Square Propagation (RMS-Prop), Mini-Batch Gradient Descent, Gradient Descent, Stochastic Gradient Descent, etcetera. are some optimizers that can help a model to increase its performance.

Adaptive Moment Estimation (Adam)

Adam, or adaptive moment estimation, is a well-known, straightforward stochastic gradient descent optimizer. In the disciplines of computer vision and natural language processing, it is also well-liked. More specifically, adam is an algorithm that enables the computation of adaptive learning rates for each parameter. Compared to RMS-Prop and AdaDelta, adam will be easier to apply in our scenario because it is more computationally efficient, needs less memory, is well-suited to problems with large amounts of data and parameters, and can produce good results quickly. Additionally, the starting decay rates provided in and affect the optimizer's learning rate. As a result, in our experiment, we utilized Adam to improve the pre-trained VGG16 and VGG19 with a learning rate of and the suggested CNN-XGBoost model with a minimum learning rate of 0.00001.

4.7.5 Activation Functions

An example of a sort of function that helps the neural network understand complex patterns in input and determine its output is activation functions. These functions help determine which parameter values should be passed on to the following tier of levels and are usually located near the conclusion of a layer. These functions also increase a model's output, accuracy, and computational effectiveness. We have used LeakyReLU in the pre-trained model such as VGG16 and VGG19 and in our proposed CNN-XGBoost model we used ReLU as an activation function. The equation of ReLU function is

$$ReLU(z_i) = max(0, z_i) \tag{4.19}$$

Leaky Rectified Linear Unit (LeakyReLU)

A piecewise linear function called the ReLU, or rectified linear activation function, outputs the input directly if it is positive and zeros otherwise. ReLU and LeakyReLU may be distinguished and explained using the straightforward graph shown below

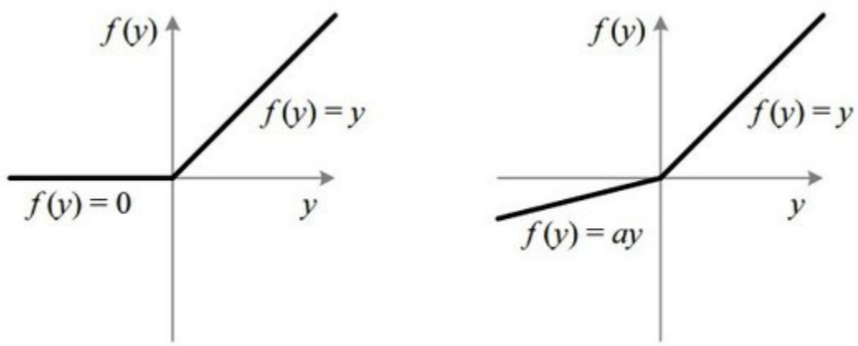


Figure 4.13: ReLU and LeakyReLU activation functions

The Leaky Rectified Linear Unit, or LeakyReLU, is totally derived from ReLU, with the exception that it has a very small slope for negative values as opposed to a flat slope. Because they expedite training and solve the dead ReLU problem that may arise if we utilized ReLU instead of LeakyReLU.

4.7.6 Softmax

Softmax activation outperforms sigmoid activation when it comes to multi-class classification. For multi-class problems with categorization, the sigmoid function shouldn't be utilized since the odds are uncertain and do not account for the probabilities of the other classes.

4.7.7 Summary of Hyperparameter

Table 4.1: Summary of Hyperparameter

Hyperparameter	**value**
Learning Rate	0.00001
Batch Size	32
Number of Epochs	100
Optimizer	Adam
Data Augmentation	padding=same
Loss Function	CrossEntropyLoss

4.7.8 Loss Functions

The difference between the predicted result and the result produced by any particular model is effectively quantified by loss functions, from which gradients may be formed to change the weights for the following layer. There are many distinct loss functions, such as mean squared error loss, mean absolute error loss, binary cross-entropy loss, hinge loss, squared hinge loss, multi-class cross-entropy loss, categorical cross-entropy loss, sparse categorical cross-entropy loss etc., each having a particular functional purpose. In our research work as we are doing multiclass classification so we used categorical cross-entropy loss and sparse categorical cross-entropy loss.

Categorical Cross Entropy

Categorical cross-entropy is a well-known loss function in issues of multi-class classification. Additionally, one-hot encoding of the labels makes categorical cross-entropy most effective. Our use of one-hot encoding to represent our string labels numerically was briefly covered in the Data Pre-processing section. As a result, we employed categorical cross-entropy as a loss function in the VGG16 and VGG19 pre-trained emotion detection models

$$L_{ce} = -\sum T_i log(S_i) \tag{4.20}$$

where, T_i is the one hot encoded truth table and S_i is the softmax probabilities.

Sparse Categorical Cross Entropy

Sparse categorical cross-entropy is a well-known loss function in issues of multi-class classification. Truth flags in sparse categorical cross-entropy generally numeric stored, such as, {1}, {2}, {3}, {4} for 4-class problem. Both the categorical and sparse cross-entropy equations are same the major difference is in sparse categorical cross-entropy the truth table is integer encoded and in categorical cross-entropy the truth table is one-hot encoded. We used this sparse categorical cross-entropy in our proposed CNN-XGBoost model.

Chapter 5

Results and Discussion

5.1 Performance Metrics

The evaluation of the trained model's effectiveness has been discussed in this chapter. Accuracy, F1 score, precision, and recall are some of the main standards we've defined for measuring the performance of our trained models. With the use of a confusion matrix, we were able to see the classification results.

5.1.1 Confusion Matrix

The prediction result is summarized in the confusion matrix together with the count values of the right and wrong guesses. The weakness of depending entirely on classification accuracy is overcome by the breakdown. The four metrics in the confusion matrix are True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

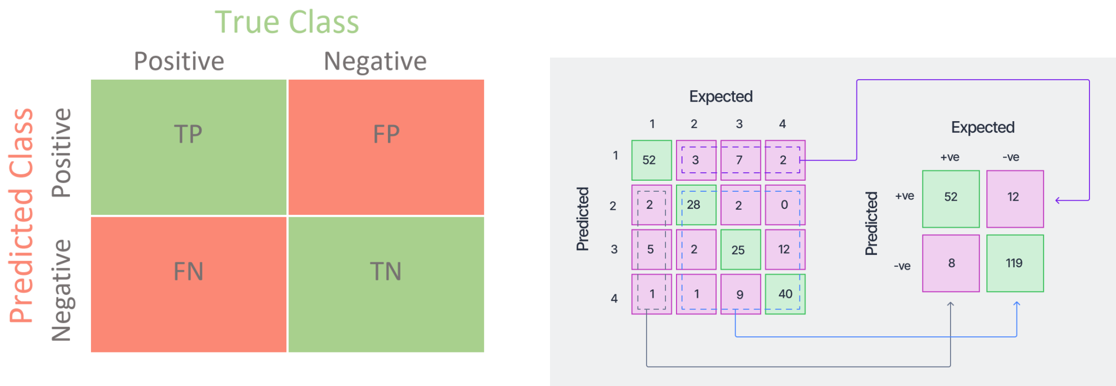


Figure 5.1: Metric of Confusion Matrix example

5.1.2 Recall

Recall, a metric for comprehensiveness, is a way to quantify a classifier's accuracy. False negatives are less common when memory is strong, but they are more frequent when recall is poor. Since accuracy decreases as the size of the data point increases, increasing recall might occasionally result in less precision.

$$recall = \frac{T_P}{T_P + F_N} \tag{5.1}$$

5.1.3 Precision

Precision is measured by the proportion of properly recognized samples to all samples that were correctly or incorrectly classed as positive. The model's precision is a metric for how effectively it can determine whether a sample is positive or negative. Less accuracy leads to more of these inaccurate readings, whereas more precision implies fewer false positives.

$$precision = \frac{T_P}{T_P + F_P} \tag{5.2}$$

5.1.4 F1 Score

The F1 Score measures the best trade-off between a classification model's accuracy and its capacity to accurately identify instances of value.

$$F1Score = 2 \times \frac{precision \times recall}{precision + recall} \tag{5.3}$$

5.1.5 Accuracy

Accuracy enables the accurate measurement of both good and negative events.

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \times 100 \tag{5.4}$$

5.2 Performance Evaluation of Spectrogram Feature Based Classification

5.2.1 Proposed CNN-XGBoost model performance

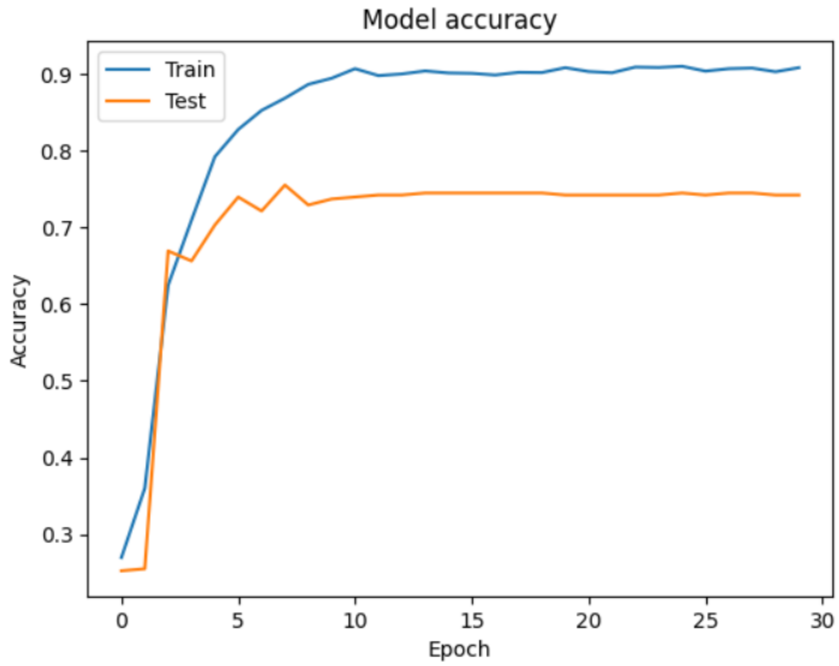


Figure 5.2: Line graph of CNN-XGBoost model accuracy

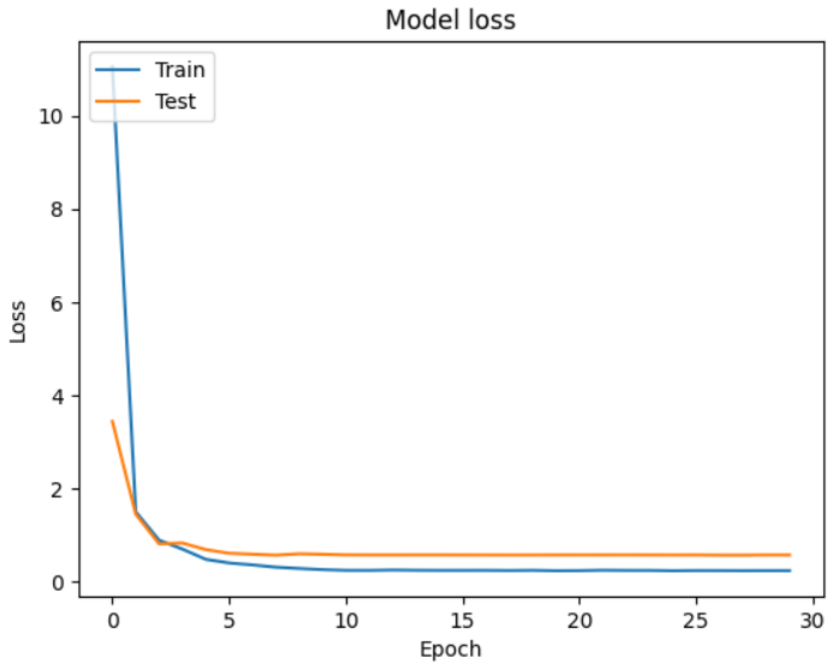


Figure 5.3: Line graph of CNN-XGBoost model loss

For our proposed CNN-XGBoost model we have initiated with 30 epochs. From the initial epoch the training accuracy of the dataset is gradually increasing and it reaches 99.94% at the last epoch. On the other hand, for the test dataset the accuracy is clicked the highest among all the experimented model about 79.43%. Which is better than our other compared fusion model CNN+Swin Transformer. On the other hand, for our proposed CNN-XGBoost model, the loss graph is gradually decreasing.

Figure 5.4: Confusion matrix graph of CNN-XGBoost model

When it comes to categorizing data, the effectiveness of a classification approach is displayed in the form of a confusion matrix, which demonstrates how effectively a categorization system performs in actual use. The effectiveness of a classification system is demonstrated and condensed in the confusion matrix.

Confusion matrix given above, observing that the Y-axis indicated the real labels, and the X-axis indicated the predicted labels by proposed CNN-XGBoost model. This model correctly predict 80 classes, 77 classes, 71 classes and 77 classes of fear, happy, neutral and sad emotions respectively.

5.2.2 Custom CNN model performance

For our custom CNN model we have initiated with 50 epochs. From the initial epoch the training accuracy of the dataset is gradually increasing and it reaches 79.94% at the last epoch. On the other hand, for the test dataset the accuracy is clicked about

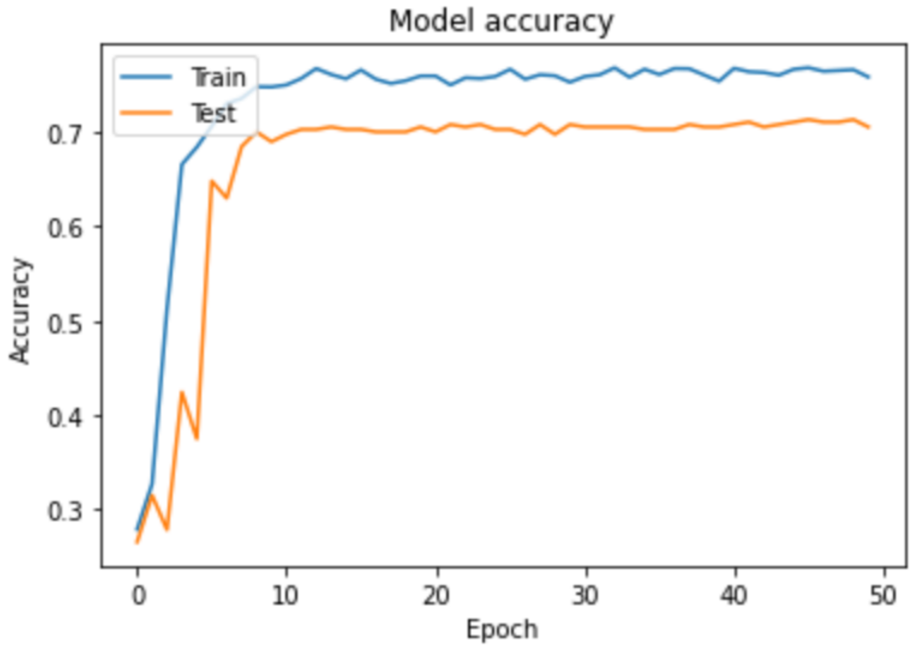


Figure 5.5: Line graph of Custom CNN model accuracy

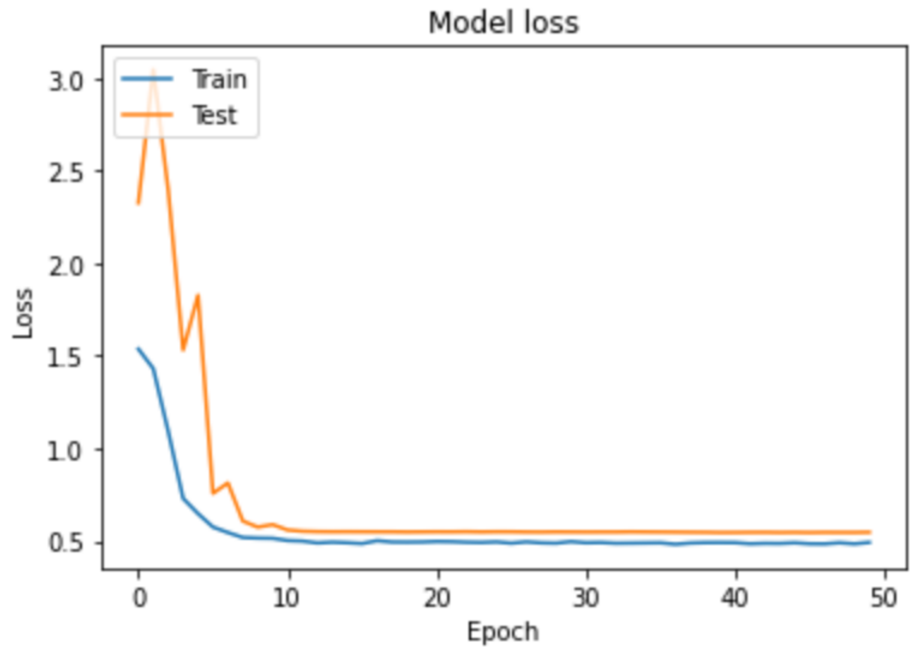


Figure 5.6: Line graph of Custom CNN model loss

72.66%. Which is the second best in our compared models. On the other hand, for our Custom CNN model, the loss graph is gradually decreasing.
When it comes to categorizing data, the effectiveness of a classification approach is displayed in the form of a confusion matrix, which demonstrates how effectively a categorization system performs in actual use. The effectiveness of a classification system is demonstrated and condensed in the confusion matrix.

Confusion matrix given above, observing that the Y-axis indicated the real labels,

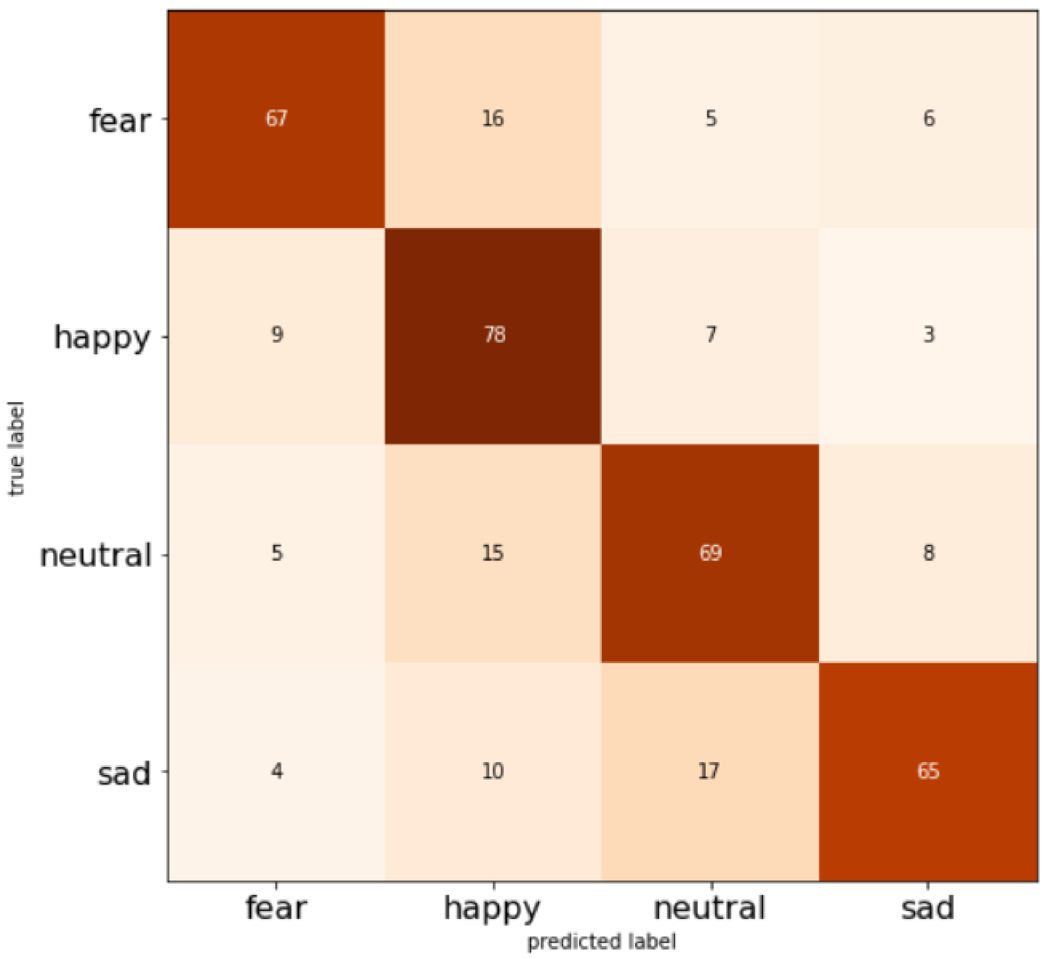


Figure 5.7: Confusion matrix graph of Custom CNN model

and the X-axis indicated the predicted labels by our Custom CNN model. This model correctly predict 67 classes, 78 classes, 69 classes and 65 classes of fear, happy, neutral and sad emotions respectively.

5.2.3 CNN+Swin Transformer model performance

For our another fusion CNN+Swin Transformer model we have initiated with 50 epochs. From the initial epoch the training accuracy of the dataset is presenting a scattered line and it reaches just only 41.94% at the last epoch. On the other hand, for the test dataset the accuracy is clicked the second lowest among all the experimented model about 40.00%.On the other hand, For CNN+Swin Transformer model, the loss graph is very scattered compared to our main proposed CNN-XGBoost model.

When it comes to categorizing data, the effectiveness of a classification approach is displayed in the form of a confusion matrix, which demonstrates how effectively a categorization system performs in actual use. The effectiveness of a classification system is demonstrated and condensed in the confusion matrix.

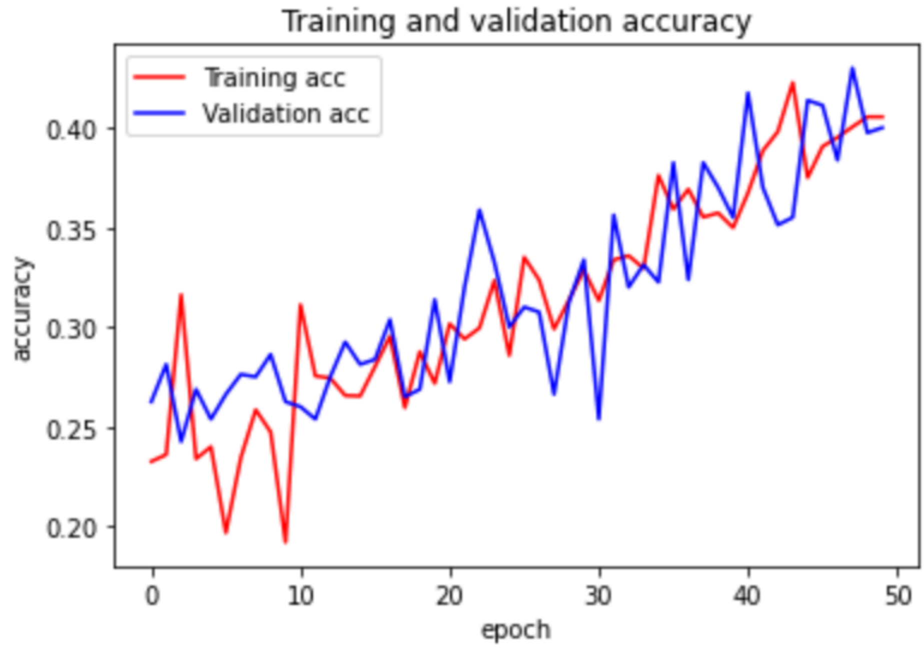


Figure 5.8: Line graph of CNN+Swin Transformer model accuracy

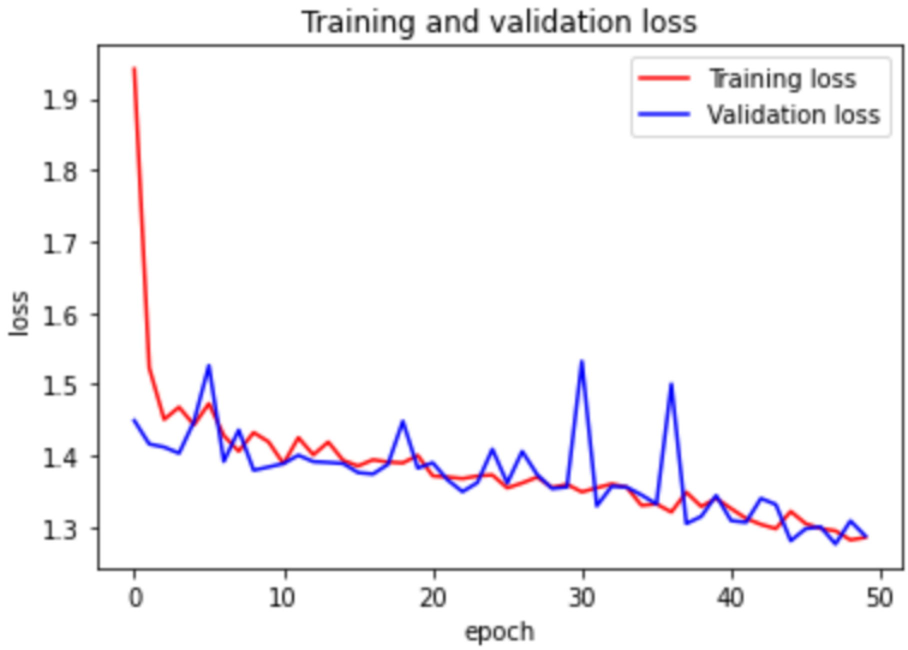


Figure 5.9: Line graph of CNN+Swin Transformer model loss

5.3 Performance Evaluation of Signal Feature Based Classification

5.3.1 VGG19 performance

For pre-trained VGG19 model we have initiated with 100 epochs. From the initial epoch the training accuracy of the dataset is gradually increasing and it reaches 95.93% at the last epoch. On the other hand, for the test dataset the accuracy

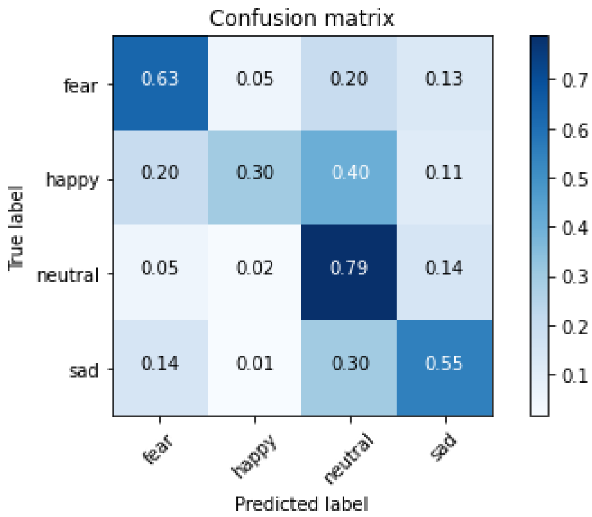


Figure 5.10: Confusion matrix graph of CNN+Swin Transformer model

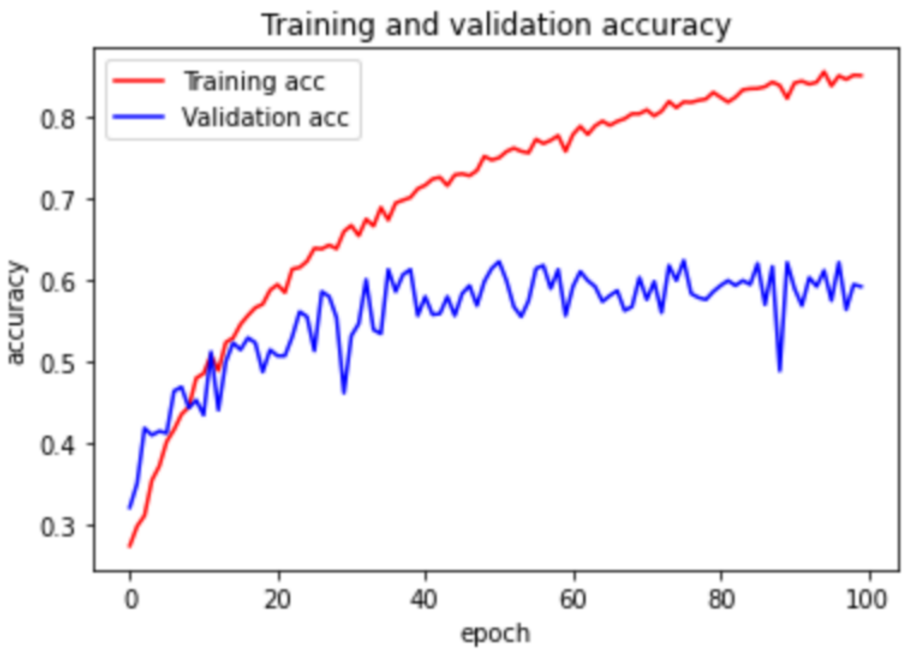


Figure 5.11: Line graph of VGG19 accuracy

is clicked about 63.66%. On the other hand, For VGG19, the loss graph is very scattered compared to our main proposed CNN-XGBoost model.

When it comes to categorizing data, the effectiveness of a classification approach is displayed in the form of a confusion matrix, which demonstrates how effectively a categorization system performs in actual use. The effectiveness of a classification system is demonstrated and condensed in the confusion matrix.

Confusion matrix given above, observing that the Y-axis indicated the real labels, and the X-axis indicated the predicted labels by pre-trained VGG19 model. This model correctly predict 127 classes, 60 classes, 160 classes and 111 classes of fear, happy, neutral and sad emotions respectively.

Figure 5.12: Line graph of VGG19 loss

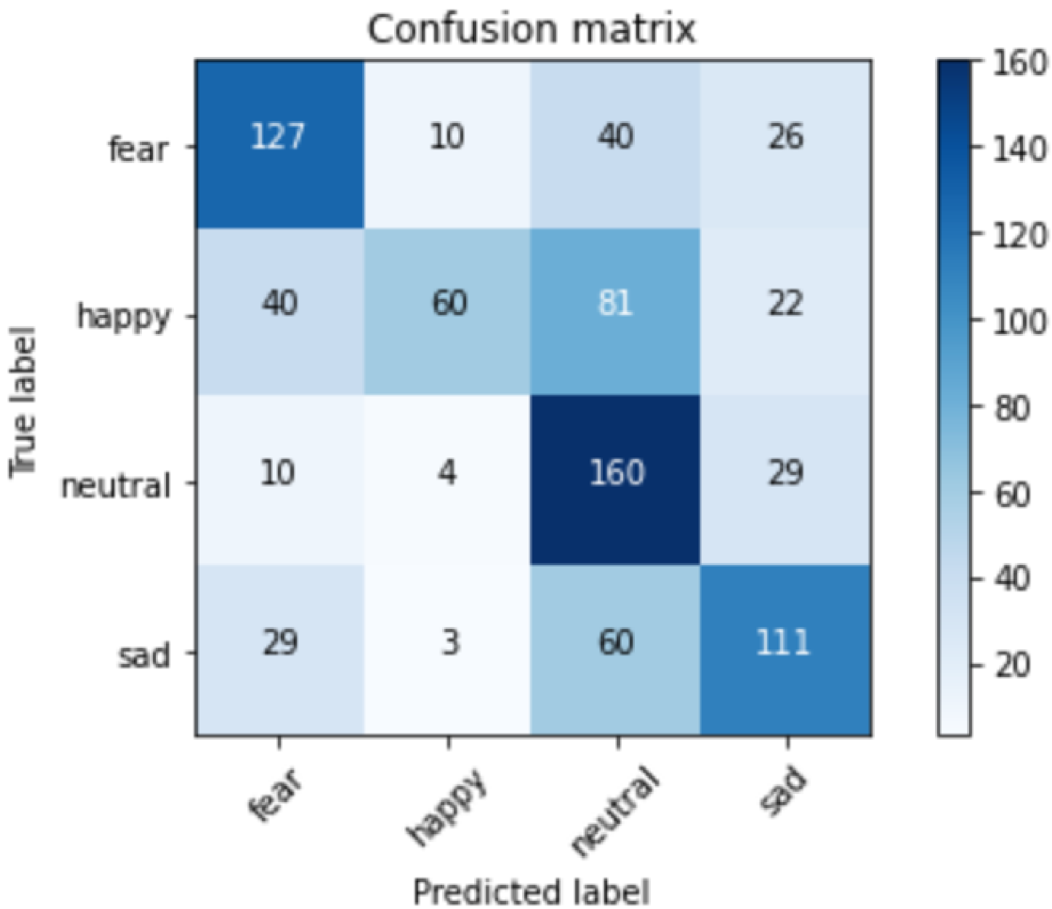


Figure 5.13: Confusion matrix VGG19

5.3.2 VGG16 performance

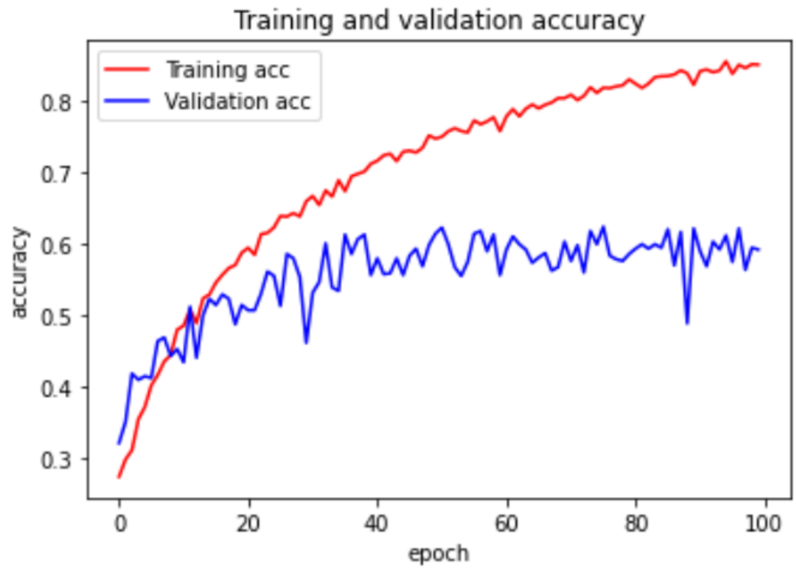


Figure 5.14: Line graph of VGG16 accuracy

Figure 5.15: Line graph of VGG16 loss

For pre-trained VGG16 model we have initiated with 100 epochs. From the initial epoch the training accuracy of the dataset is gradually increasing and it reaches 94.93% at the last epoch. On the other hand, for the test dataset the accuracy is clicked about 60.01%. On the other hand, For VGG16, the loss graph is very scattered compared to our main proposed CNN-XGBoost model.

When it comes to categorizing data, the effectiveness of a classification approach is displayed in the form of a confusion matrix, which demonstrates how effectively a categorization system performs in actual use. The effectiveness of a classification

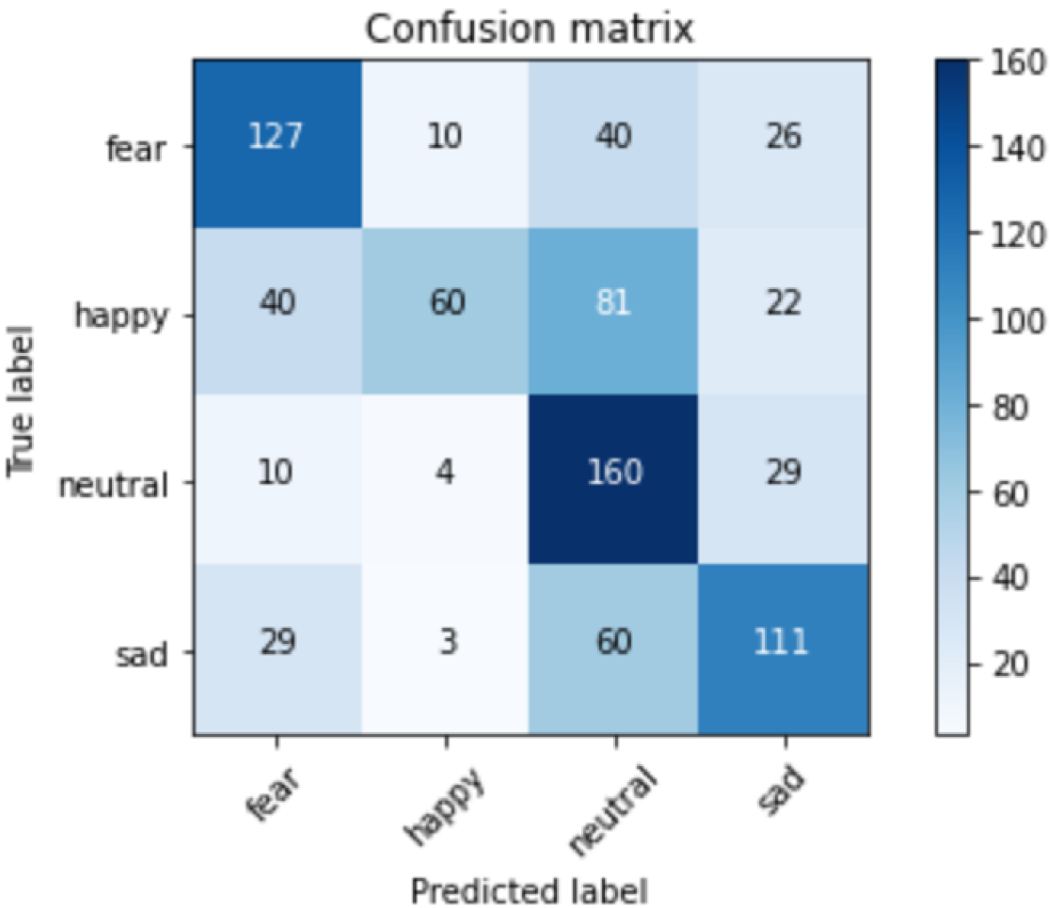


Figure 5.16: Confusion matrix VGG16

system is demonstrated and condensed in the confusion matrix.

Confusion matrix given above, observing that the Y-axis indicated the real labels, and the X-axis indicated the predicted labels by pre-trained VGG16 model. This model correctly predict 127 classes, 60 classes, 160 classes and 111 classes of fear, happy, neutral and sad emotions respectively.

5.3.3 LSTM performance

For LSTM model we have initiated with 100 epochs. From the initial epoch the training accuracy of the dataset is gradually increasing and it reaches 97.93% at the last epoch. On the other hand, for the test dataset the accuracy is clicked about 59.01%. This model we used for just experimental basis as how it works some works to be needed to re-burnish this seems overfitted with the data.

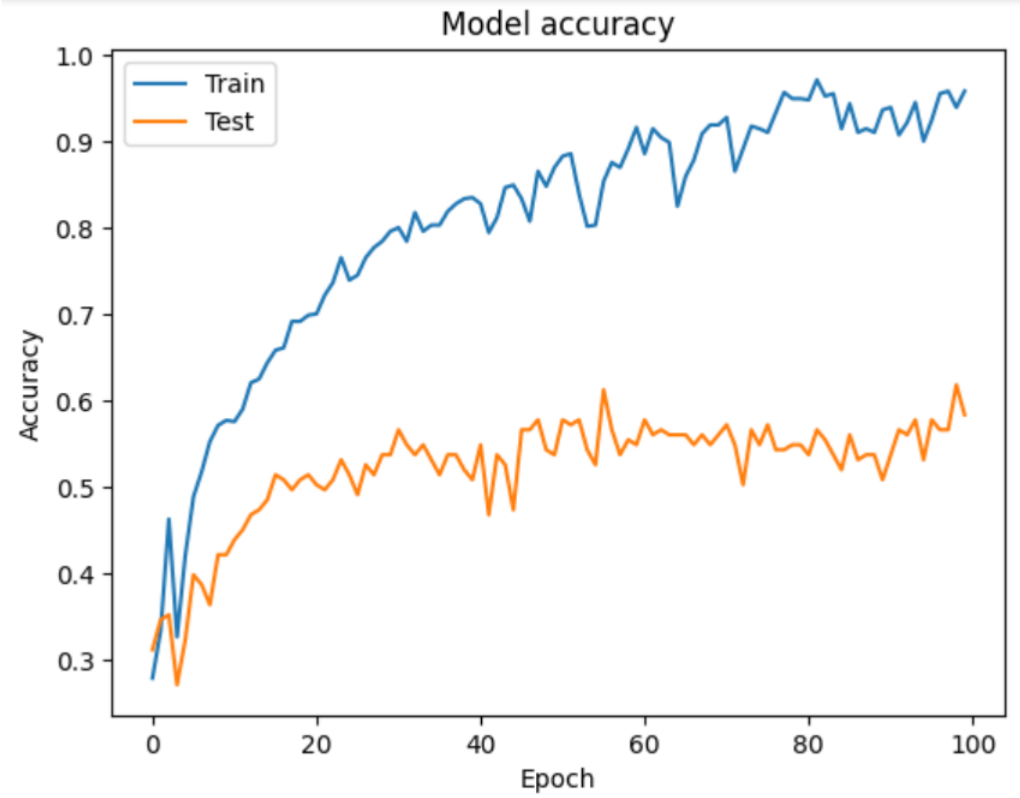


Figure 5.17: Line graph of LSTM accuracy

5.4 Classification Reports and Comparisons

The experiment includes testing six different variants of the model. We found that our proposed CNN-XGBoost fusion approach can classify highest number of correct emotions with the context of accuracy.

Table 5.1: Performance Comparison of Different Models

Model	Accuracy	Precision	Recall	F1 Score
Custom CNN	72.66%	70%	71%	71%
VGG19	63.66%	59%	55%	59%
VGG16	60.01%	59%	55%	57%
CNN+Swin Transformer	40.0%	65%	53%	55%
LSTM	59.01%	65%	60%	63%
CNN-XGBoost [Proposed]	79.43%	78%	78%	78%

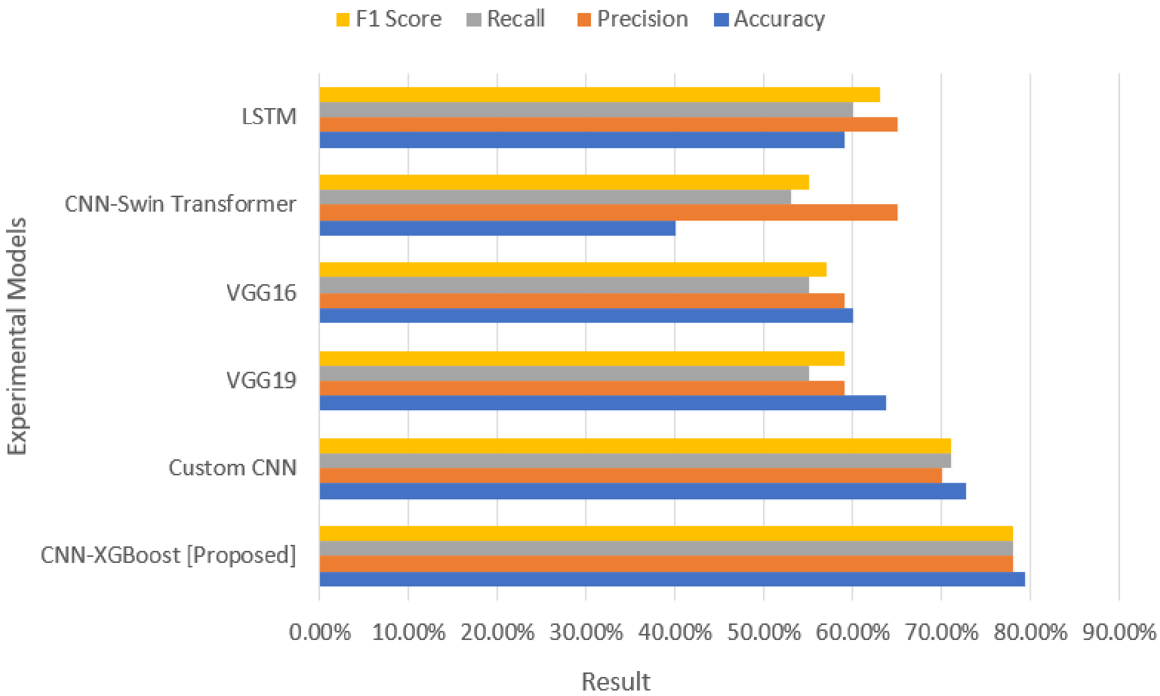


Figure 5.18: Clustered Bar Performance Comparison graph of Different Models.

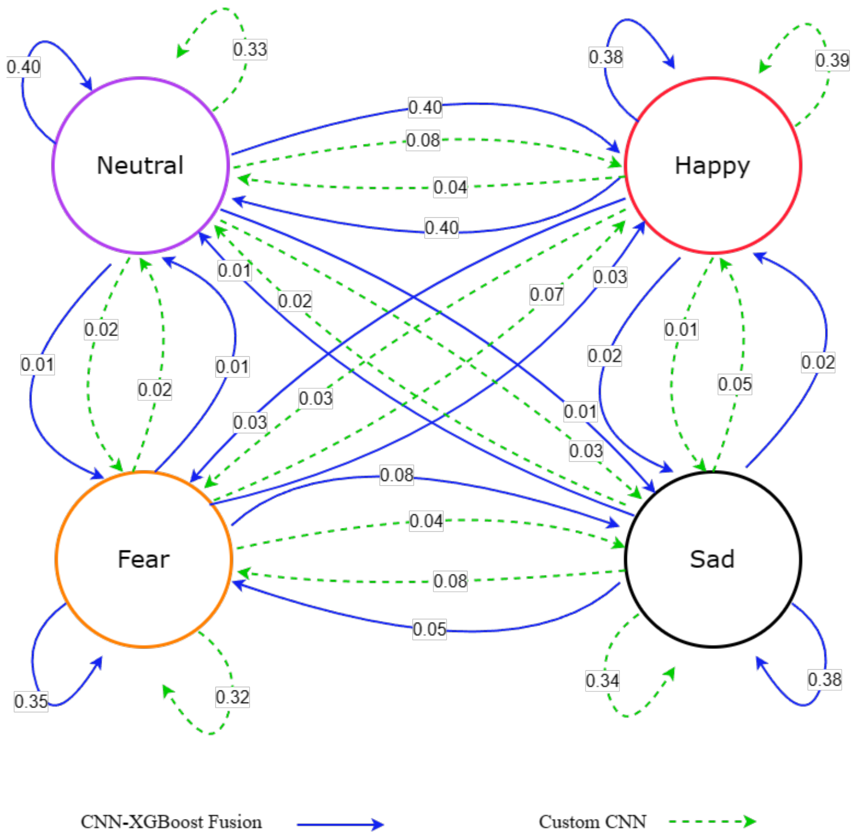


Figure 5.19: Normalize Confusion graph of CNN-XGBoost and Custom CNN model comparison, it demonstrates how its enhancing capabilities for recognizing emotions act.

Properly Classified Spectrogram Images

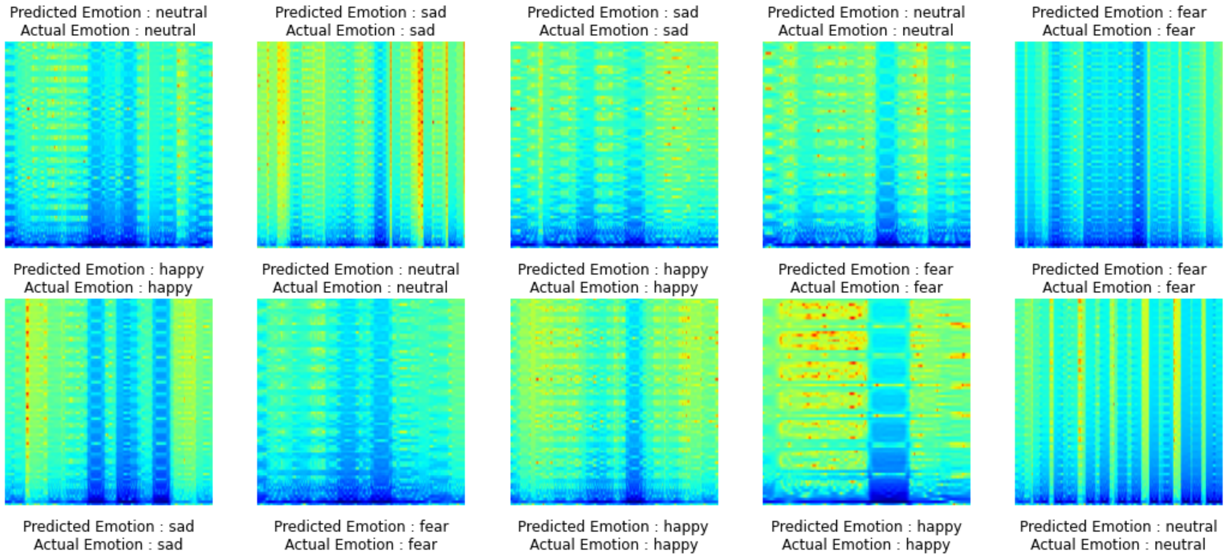


Figure 5.20: Properly Classified Spectrogram Images

Misclassified Spectrogram Images

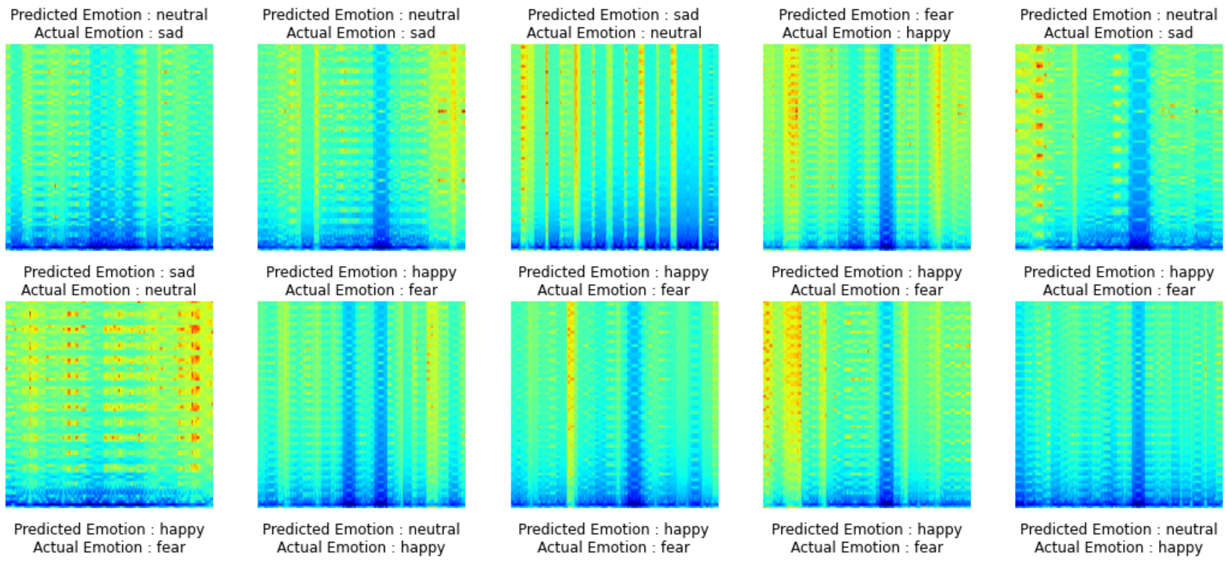


Figure 5.21: Misclassified Spectrogram Images

Performance Comparison with Related Work Methods

Table 5.2: Performance Comparison with Related Work Methods

Related Paper	Dataset	Procedure	Validity
Zheng et al. [21]	SEED-IV [4]	Statistical features, SVM	70%
Fan et al. [10]	AFEW 6.0 [45]	Statistical features, 3DCNN-RNN	59.02%
Yao et al. [11]	Emotiw [46]	HoloNet	57.84%
Zhang et al. [12]	RECOLA [47]	SSL	76.05%
Kim et al. [13]	SAVEE [48]	Data Driven	65.60%
Proposed method	SEED-IV [4]	Spectrogram features, CNN-XGBoost	79.43%

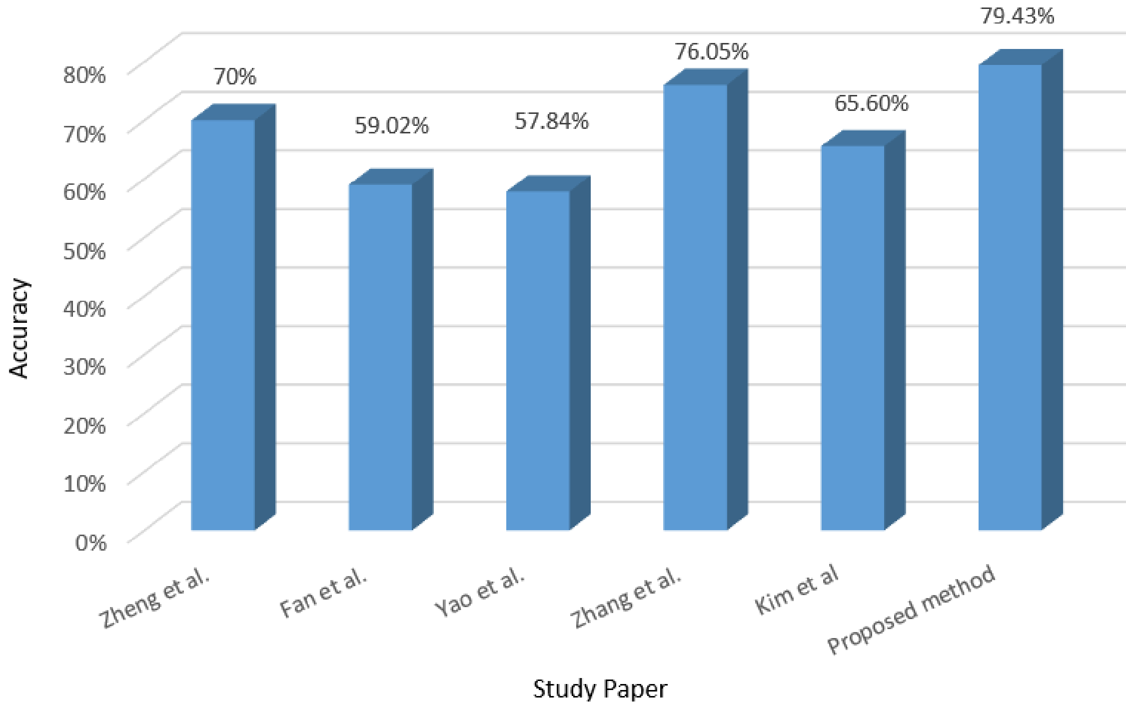


Figure 5.22: Clustered Column Performance Comparison graph of Related Work Methods.

Our benchmark was 70.0% and by proposing a new fusion approach CNN-XGBoost by using Spectrogram image analysis we have successfully break our benchmark and gain 79.43% of accuracy.

Chapter 6

Conclusion

As a conclusion to the study, we will explore the limitations of our work and potential results. We will also talk about prospective directions for our study.

6.1 Final Remarks

In conclusion, the proposed fusion-based CNN-XGBoost model for multiclass emotion classification using EEG spectrogram images outperforms the comparison models such as VGG16, VGG19, Swin Transformer etc. The accuracy of the VGG16 model's test dataset varied between 57% and 60%, despite having been trained for 100 epochs and reaching a training accuracy of 89.24% in the final epoch. In comparison, our suggested model performed better. Compared to our primary CNN-XGBoost model, the VGG16 loss graph showed substantial dispersion, indicating less stable training. This shows that our fusion-based model offers learning that is more robust and reliable.The confusion matrix graph of VGG16 additionally gives insight into the potency of the classification strategy. It accurately predicted 111 cases of sadness, 60 cases of happiness, 127 cases of fear, and 160 cases of neutral mood. However, as was already indicated, our suggested CNN-XGBoost fusion model produced much better outcomes. These results reinforce the proposed fusion-based model's superiority to VGG16 and stress its accuracy and efficacy in categorizing multiclass emotions. A more sophisticated and trustworthy system for emotion recognition is made possible by using EEG spectrogram images in conjunction with the strength of CNNs and XGBoost algorithms.Overall, by showcasing the improved functionality of the CNN-XGBoost fusion-based model, this research advances the field of emotion recognition. When compared to models like CNN+Swin Transformer and VGG16, the model shows improved accuracy and robustness. The results of the confusion matrix study provide additional support for the usefulness of our suggested paradigm. It successfully categorizes a large number of examples for each emotion category, demonstrating its accuracy in recognizing fear, happiness, neutrality, and sadness. These findings lend credence to the idea that a fusion-based method that makes use of EEG spectrogram pictures can identify emotions with greater accuracy and dependability. As we worked on SEED-IV dataset [4] the paper of the dataset Zheng et al. [21] had 70.0% accuracy with the statistical features and with SVM classifier that was oir benchmark. In this research we tried to work with another time domain frequency feature which is spectrogram image analysis and by using a fusion approach of CNN-XGBoost we have successfully break our benchmark and achieve about 79.43% of accuracy.

6.2 Limitations

It is important to acknowledge the limitations of this study. One must prepare for encountering a lot of procedural difficulties during the study process. The following are some difficulties over the work.

6.2.1 Access

This serves as an essential example of the limitations imposed on the dissertation study. Accessing resources that are crucial to our study, such as papers, people, or other resources, has been difficult for us as we have been writing the dissertation.

6.2.2 Data Aggregation

Real EEG data could not be used because our university did not have an EEG headset; therefore, we had to rely on datasets that were made accessible to the public. The generality of the outcomes may be impacted by this. Additionally, even though different feature extraction strategies were investigated, it's possible that other approaches were overlooked. Since only a small number of comparison models were employed in the evaluation. As the high computational power also was the big barrier for us as we worked with image analysis so it's possible that other cutting-edge models or ensemble methodologies would produce different outcomes. Despite these drawbacks, this research lays the groundwork for subsequent investigations that will overcome these issues and enhance the field of multiclass emotion classification through the use of fusion-based methods and EEG spectrogram analysis.

6.2.3 Time restraint

The quantity of time available could be seen as a constraint in some situations, although this will much depend on the kind of study being done. For us, completing the study within the allotted time was essential. Despite this, we were still able to properly finish the research and utilize the time we had.

6.3 Future Works

The development of future systems for human-computer interaction will be significantly impacted by the suggested model's improved accuracy, stability, and classification efficacy. These systems can provide personalized services, enhance user experiences, and enable more effective communication between humans and machines by precisely identifying and comprehending human emotions. In future, we would extend this work using multimodal strategy to contribute into the sector of neuromarketing.

Bibliography

[1] M. H. Siddiqi, M. G. R. Alam, C. S. Hong, A. M. Khan, and H. Choo, "A novel maximum entropy markov model for human facial expression recognition," *PLOS ONE*, vol. 11, no. 9, pp. 1–23, Sep. 2016. DOI: 10.1371/journal.pone.0162702. [Online]. Available: https://doi.org/10.1371/journal.pone.0162702.

[2] S. Katsigiannis and N. Ramzan, "Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 98–107, 2018. DOI: 10.1109/JBHI.2017.2688239.

[3] P. Murugappan, M. Juhari, R. Nagarajan, and S. Yaacob, "An investigation on visual and audiovisual stimulus based emotion recognition using eeg," *International Journal of Medical Engineering and Informatics*, vol. 1, Jan. 2009. DOI: 10.1504/IJMEI.2009.022645.

[4] *Seed-iv dataset.* [Online]. Available: https://bcmi.sjtu.edu.cn/home/seed/seed-iv.html.

[5] M. Royo-Vela and Á. Varga, "Unveiling neuromarketing and its research methodology," *Encyclopedia*, vol. 2, no. 2, pp. 729–751, 2022, ISSN: 2673-8392. DOI: 10.3390/encyclopedia2020051. [Online]. Available: https://www.mdpi.com/2673-8392/2/2/51.

[6] M. Zhao, F. Adib, and D. Katabi, "Emotion recognition using wireless signals," in *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '16, New York City, New York: Association for Computing Machinery, 2016, pp. 95–108, ISBN: 9781450342261. DOI: 10.1145/2973750.2973762. [Online]. Available: https://doi.org/10.1145/2973750.2973762.

[7] J. Atkinson and D. Campos, "Improving bci-based emotion recognition by combining eeg feature selection and kernel classifiers," *Expert Syst. Appl.*, vol. 47, no. C, pp. 35–41, Apr. 2016, ISSN: 0957-4174. DOI: 10.1016/j.eswa.2015.10.049. [Online]. Available: https://doi.org/10.1016/j.eswa.2015.10.049.

[8] Y. Li, Y. Jiang, D. Tian, L. Hu, H. Lu, and Z. Yuan, "Ai-enabled emotion communication," *IEEE Network*, vol. 33, no. 6, pp. 15–21, 2019. DOI: 10.1109/MNET.001.1900070.

[9] Y. Lei, S. Yang, X. Wang, and L. Xie, "Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 853–864, 2022. DOI: 10.1109/TASLP.2022.3145293.

[10] Y. Fan, X. Lu, D. Li, and Y. Liu, “Video-based emotion recognition using cnn-rnn and c3d hybrid networks,” Nov. 2016. DOI: 10.1145/2993148.2997632.

[11] A. Yao, D. Cai, P. Hu, S. Wang, L. Sha, and Y. Chen, “Holonet: Towards robust emotion recognition in the wild,” in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ser. ICMI ’16, Tokyo, Japan: Association for Computing Machinery, 2016, pp. 472–478, ISBN: 9781450345569. DOI: 10.1145/2993148.2997639. [Online]. Available: https://doi.org/10.1145/2993148.2997639.

[12] Z. Zhang, F. Ringeval, B. Dong, E. Coutinho, E. Marchi, and B. Schuller, “Enhanced semi-supervised learning for multimodal emotion recognition,” Mar. 2016, pp. 5185–5189. DOI: 10.1109/ICASSP.2016.7472666.

[13] Y. Kim and E. M. Provost, “Emotion recognition during speech using dynamics of multiple regions of the face,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 12, pp. 1–23, Oct. 2015. DOI: 10.1145/2808204.

[14] W.-L. Zheng and B.-L. Lu, “Personalizing eeg-based affective models with transfer learning,” in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, ser. IJCAI’16, New York, New York, USA: AAAI Press, 2016, pp. 2732–2738, ISBN: 9781577357704. [Online]. Available: https://www.ijcai.org/Proceedings/16/Papers/388.pdf.

[15] H. C. van Cuylenburg and T. Ginige, “Emotion guru: A smart emotion tracking application with ai conversational agent for exploring and preventing depression,” in *2021 International Conference on UK-China Emerging Technologies (UCET)*, 2021, pp. 1–6. DOI: 10.1109/UCET54125.2021.9674993.

[16] X. Wu, W.-L. Zheng, Z. Li, and B.-L. Lu, “Investigating eeg-based functional connectivity patterns for multimodal emotion recognition,” *Journal of Neural Engineering*, vol. 19, no. 1, p. 016 012, Jan. 2022. DOI: 10.1088/1741-2552/ac49a7. [Online]. Available: https://dx.doi.org/10.1088/1741-2552/ac49a7.

[17] H. Zeng, X. Wang, A. Wu, *et al.*, “Emoco: Visual analysis of emotion coherence in presentation videos,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 927–937, 2020. DOI: 10.1109/TVCG.2019.2934656.

[18] C.-J. Yang, N. Fahier, W.-C. Li, and W.-C. Fang, “A convolution neural network based emotion recognition system using multimodal physiological signals,” in *2020 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)*, 2020, pp. 1–2. DOI: 10.1109/ICCE-Taiwan49838.2020.9258341.

[19] Y. Mouammine and H. Azdimousa, “Using neuromarketing and ai to collect and analyse consumer’s emotion: Literature review and perspectives,” Dec. 2019.

[20] A. Al-Nafjan, “Feature selection of EEG signals in neuromarketing,” 2022. DOI: 10.7717/peerj-cs.944.

[21] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, “Emotionmeter: A multimodal framework for recognizing human emotions,” *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 1110–1122, 2019. DOI: 10.1109/TCYB.2018.2797176.

[22] M. R. Islam, M. A. Moni, M. M. Islam, *et al.*, "Emotion recognition from eeg signal focusing on deep learning and shallow learning techniques," *IEEE Access*, vol. 9, pp. 94 601–94 624, 2021. DOI: 10.1109/access.2021.3091487.

[23] M. S. Khan, N. Salsabil, M. G. Alam, M. A. Dewan, and M. Z. Uddin, "Cnn-xgboost fusion-based affective state recognition using eeg spectrogram image analysis," *Scientific Reports*, vol. 12, no. 1, 2022. DOI: 10.1038/s41598-022-18257-x.

[24] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015. DOI: 10.1109/TAMD.2015.2431497.

[25] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "Emotionmeter: A multimodal framework for recognizing human emotions," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 1110–1122, 2019. DOI: 10.1109/TCYB.2018.2797176.

[26] Y. Lu, W.-L. Zheng, B. Li, and B.-L. Lu, "Combining eye movements and eeg to enhance emotion recognition," in *International Joint Conference on Artificial Intelligence*, 2015.

[27] W. Liu, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal deep learning," vol. 9948, Oct. 2016, ISBN: 978-3-319-46671-2. DOI: 10.1007/978-3-319-46672-9_58.

[28] H. Tang, W. Liu, W.-L. Zheng, and B.-L. Lu, "Multimodal emotion recognition using deep neural networks," Oct. 2017, pp. 811–819, ISBN: 978-3-319-70092-2. DOI: 10.1007/978-3-319-70093-9_86.

[29] P. Murugappan, "Human emotion recognition through short time electroencephalogram (eeg) signals using fast fourier transform (fft)," Jun. 2013. DOI: 10.1109/CSPA.2013.6530058.

[30] Y. Chavhan, M. Dhore, and Y. Pallavi, "Speech emotion recognition using support vector machines," *International Journal of Computer Applications*, vol. 1, Feb. 2010. DOI: 10.1007/978-3-642-21402-8_35.

[31] P. Murugappan, W. Khairunizam, S. Yaacob, and J. Selvaraj, "Electrocardiogram-based emotion recognition system using empirical mode decomposition and discrete fourier transform," *Expert Systems*, vol. 31, Mar. 2013. DOI: 10.1111/exsy.12014.

[32] D. Acharya, A. Billimoria, N. Srivastava, S. Goel, and A. Bhardwaj, "Emotion recognition using fourier transform and genetic programming," *Applied Acoustics*, vol. 164, p. 107 260, Jul. 2020. DOI: 10.1016/j.apacoust.2020.107260.

[33] S. Jirayucharoensak, S. Pan-ngum, and P. Israsena, "Eeg-based emotion recognition using deep learning network with principal component based covariate shift adaptation," *TheScientificWorldJournal*, vol. 2014, p. 627 892, Sep. 2014. DOI: 10.1155/2014/627892.

[34] A. E Vijayan, "Eeg-based emotion recognition using statistical measures and auto-regressive modeling," Feb. 2015. DOI: 10.1109/CICT.2015.24.

[35] M. Hassan, M. G. R. Alam, M. Z. Uddin, S. Huda, A. Almogren, and G. Fortino, "Human emotion recognition using deep belief network architecture," *Information Fusion*, vol. 51, Nov. 2018. DOI: 10.1016/j.inffus.2018.10.009.

[36] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015. DOI: 10.1109/TAMD.2015.2431497.

[37] S. L. Y. BL; *Differential entropy feature for eeg-based vigilance estimation.* [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/24111262/.

[38] [Online]. Available: https://www.researchgate.net/profile/Bao-Liang-Lu/publication/261239872_Differential_entropy_feature_for_EEG-based_emotion_classification/links/58c6a2b2aca272e36dde6cba/Differential-entropy-feature-for-EEG-based-emotion-classification.pdf.

[39] W. Chen, M. J. Er, and S. Wu, "Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 2, pp. 458–466, 2006. DOI: 10.1109/TSMCB.2005.857353.

[40] G. Kharat and S. Dudul, "Neural network classifier for human emotion recognition from facial expressions using discrete cosine transform," Aug. 2008, pp. 653–658, ISBN: 978-0-7695-3267-7. DOI: 10.1109/ICETET.2008.22.

[41] H. Ali, H. M, S. Yaacob, and A. Adom, "Facial emotion recognition using empirical mode decomposition," *Expert Systems with Applications*, vol. 42, pp. 1261–1277, Feb. 2015. DOI: 10.1016/j.eswa.2014.08.049.

[42] Z. Hafed and M. Levine, "Face recognition using the discrete cosine transform," *International Journal of Computer Vision*, vol. 43, pp. 167–188, Jul. 2001. DOI: 10.1023/A:1011183429707.

[43] S. Tammina, "Transfer learning using vgg-16 with deep convolutional neural network for classifying images," vol. 9, Oct. 2019, p9420. DOI: 10.29322/IJSRP.9.10.2019.p9420.

[44] *Imagenet.* [Online]. Available: http://www.image-net.org.

[45] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE Multimedia*, vol. 19, pp. 34–31, Sep. 2012.

[46] A. Dhall, R. Goecke, J. Joshi, J. Hoey, and T. Gedeon, "Emotiw 2016: Video and group-level emotion recognition challenges," Oct. 2016, pp. 427–432. DOI: 10.1145/2993148.2997638.

[47] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–8. DOI: 10.1109/FG.2013.6553805.

[48] *Savee dataset.* [Online]. Available: http://kahlan.eps.surrey.ac.uk/savee/.