"Real-Time Facial Expression Recognition with Bengali Audio Feedback: Bridging Communication Gaps"

by

Arham Ahmed Jobayer
20301216
Arnab Sarker Sangit
19201100
Sharthak Das
23241033

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

Department of Computer Science and Engineering
Brac University
January 2025

Declaration

It is hereby declared that

1. The thesis submitted is my/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:

Arham Ahmed Jobayer

20301216

Arnab Sarker Sangit

19201100

Sharthak Das

23241033

Approval

The thesis/project titled "Real-Time Facial Expression Recognition with Bengali Audio Feedback: Bridging Communication Gaps" submitted by

1. Arham Ahmed Jobayer (20301216)
2. Arnab Sarker Sangit (19201100)
3. Sharthak Das (23241033)

Of Fall, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January, 2025.

Examining Committee:

Supervisor:
(Member)

Mr. Moin Mostakim

Senior Lecturer
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Professor
Department of Computer Science and Engineering
Brac University

Abstract

This study investigates the efficacy of deep learning models in facial expression recognition while incorporating Bengali audio feedback. We utilize a meticulously curated dataset comprising diverse facial images depicting individuals expressing various emotions, annotated with corresponding Bengali audio descriptions. Each image is labeled with the emotion it represents, and the dataset includes metadata such as age, gender, and cultural context. We explore the performance of convolutional neural networks (CNNs), recurrent neural networks (RNNs), and hybrid models in recognizing facial expressions from images and associating them with Bengali audio feedback. Additionally, we assess the impact of data augmentation techniques on model performance. Our experiments reveal that hybrid CNN-RNN models achieve the highest accuracy in recognizing facial expressions and generating appropriate Bengali audio feedback. Furthermore, we analyze the robustness of the models across diverse demographic groups within the dataset. This study contributes to the advancement of multimodal deep learning techniques for enhancing communication experiences, particularly in contexts where Bengali audio feedback is essential

Keywords: Facial expression recognition, Bengali audio feedback, Curated dataset, Diverse facial images, Emotions, Annotations, Bengali audio descriptions, Age, Gender, Cultural context, Hybrid models, Demographic groups, Multimodal deep learnings.

Dedication (Optional)

A dedication is the expression of friendly connection or thanks by the author towards another person. It can occupy one or multiple lines depending on its importance. You can remove this page if you want.

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.
Secondly, to our co-advisor Teacher Name sir for his kind support and advice in our work. He helped us whenever we needed help.
Thirdly, Name and the whole judging panel of Conference Name. Though our paper not accepted there, all the reviews they gave helped us a lot in our later works.
And finally to our parents without their throughout sup-port it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

In our progressively interconnected world, consistent communication over dialects and societies is more pivotal than ever. This proposition, titled "Real-Time Facial Expression Acknowledgment with Bengali Sound Input: Bridging the Communication Separate," handles a basic jump by investigating the integration of profound learning for facial expression acknowledgment with Bengali sound input.. This state-of-the-art approach uses a carefully curated dataset containing various facial expressions carefully tagged with corresponding Bengali audio descriptions. Using this wealth of data, research aims to improve the accuracy and reliability of emotion recognition systems [23]. The study explores an exciting set of deep learning architectures to identify the most effective approach to this multifaceted task. Convolutional Neural Networks (CNN) are excellent at extracting complex spatial features from visual data of facial expressions, thereby detecting subtle variations in muscle movements. Recurrent Neural Networks (RNNs), on the other hand, bring their sequential data processing strengths to the table. They can effectively capture the temporal dynamics of facial expressions, which is crucial for understanding the nuances of human emotions. By combining the powers of CNN and RNN in innovative hybrid models, the study opens up possibilities for even more efficient and accurate emotion detection.. In addition, it carefully evaluates the effect of data augmentation techniques, which is the process of enriching the data set to improve the generalization of the model. In addition, the study carefully analyzes the performance of the models in different population groups to ensure inclusion of the data set [8]. The findings highlight the enormous potential of hybrid CNN-RNN models to achieve ultimate accuracy. This success paves the way for the development of multimodal deep learning techniques that can significantly transform communication experiences, especially in Bengali Dominant regions. Imagine real-time translators that not only record the spoken word, but also convey emotional delays through voice and facial expressions.. On the Top, it can revolutionize communication in educational environments where language barriers can hinder learning. For example, an AI-powered translator can facilitate real-time conversations between teachers and students who speak different languages and ensure that all students receive the same quality education regardless of their native language. This technology can improve communication in health care, business and social interactions by providing information about emotions. For example, in healthcare, emotionally intelligent AI could

help doctors better understand and respond to patients' emotional states, leading to more compassionate care. In addition, this technology could be integrated into virtual assistants and chatbots, allowing them to provide more natural and emotionally intelligent interactions with users from different backgrounds, such as a customer service bot that can detect frustration in a user's voice and better adapt its response accordingly with enhanced support. The impact extends far beyond facial expression recognition. This research promotes inclusive communication by incorporating linguistic diversity into technology. This could revolutionize communication in educational settings, healthcare environments, and any situation where people from different cultural backgrounds interact. Additionally, this technology could be integrated into virtual assistants and chatbots, enabling them to provide more natural and emotionally intelligent interactions with users from diverse backgrounds.

1.2 Research Problem

Communication is a fundamental aspect of human interaction, encompassing both verbal and non-verbal elements. Facial expressions, as one of the primary forms of non-verbal communication, play a crucial role in conveying emotions and intentions. In various contexts, such as customer service, education, and healthcare, the ability to accurately recognize and respond to facial expressions can significantly enhance interaction quality and outcomes. Despite significant advancements in facial expression recognition (FER) technologies, real-time implementation with accurate feedback remains a challenge, particularly when considering linguistic diversity and real-world applicability. Real-time FER involves the continuous analysis of facial expressions from live video feeds, requiring robust algorithms capable of processing data at high speeds without compromising accuracy. This task is inherently complex due to the dynamic nature of facial expressions, variations in lighting conditions, occlusions, and individual differences in expressive behavior. Existing systems often fall short in maintaining high performance under these varied conditions, highlighting a critical gap in current FER technologies.While many FER systems provide feedback in widely spoken languages like English, there is a notable lack of support for less commonly used languages, such as Bengali. Bengali, spoken by over 230 million people worldwide, presents unique phonetic and syntactic characteristics that necessitate specialized handling in audio feedback systems [5]. The absence of Bengali language support in FER systems limits their accessibility and effectiveness for Bengali-speaking populations, thereby perpetuating communication gaps in diverse environments.The integration of real-time Facial Expression Recognition (FER) with Bengali audio feedback serves as a pivotal step towards enhancing communication inclusivity and efficacy for Bengali-speaking users. By amalgamating these two technologies, the aim is to bridge the communication gap that might exist due to linguistic barriers, thereby ensuring a more comprehensive and effective communication solution. This integration entails the development of a system capable not only of accurately recognizing facial expressions in real-time but also delivering corresponding feedback in Bengali. Such an approach is crucial for effectively conveying the emotional and expressive nuances inherent in human communication.The overarching goal of this thesis is to design and evaluate a real-time FER system with regional based audio feedback, which revolves around addressing several key research questions.By leveraging cutting-edge technology, this thesis

endeavors to create a seamless and inclusive communication experience for Bengali-speaking individuals. Firstly, the focus lies on optimizing real-time FER for high accuracy and speed across diverse conditions. This entails exploring and implementing methodologies that can enhance the performance of FER algorithms in real-time scenarios, considering factors such as lighting variations, facial occlusions, and diverse facial expressions. Secondly, the thesis aims to identify the most effective methods for seamlessly integrating Bengali audio feedback with FER outputs. This involves devising mechanisms to synchronize facial expression recognition results with corresponding audio feedback in Bengali, ensuring a coherent and natural communication experience for users. Thirdly, attention is directed towards designing the system to accommodate the specific phonetic and syntactic characteristics of the Bengali language. This entails tailoring the system's language processing components to accurately interpret and generate Bengali audio feedback, taking into account linguistic nuances and variations inherent in the Bengali language. Finally, the thesis seeks to assess the potential impacts of such a system on communication efficacy in various contexts. This involves evaluating the system's performance and user satisfaction across different usage scenarios and demographics, thereby gauging its effectiveness in facilitating communication among Bengalispeaking individuals. This thesis proposes a novel ensemble method combining machine learning and deep learning for Facial Expression Recognition (FER) with real-time Bengali audio feedback. The FER system will be rigorously trained and validated for diverse scenarios to enhance communication for Bengali users. A Bengali text-to-speech system will be developed to convert FER outputs into natural-sounding audio, considering pronunciation and prosody. The entire system, including FER, audio generation, and user experience, will be evaluated through user studies and performance tests to ensure accuracy, responsiveness, and user satisfaction in real-world settings. This research aims to contribute to affective computing and human-computer interaction by providing an inclusive FER solution for Bengali speakers, while also offering insights into integrating linguistic diversity into future FER technologies. Capturing Facial Expression Recognition (FER) while engaging in physical activities such as walking, running, or performing various tasks presents a formidable challenge. The dynamic nature of these movements introduces complexities in identifying specific facial points or Action Units (AUs) [7], making it difficult for FER systems to accurately interpret facial expressions. This difficulty arises from the rapid changes in facial semantics and expressions during motion, requiring robust algorithms capable of discerning nuanced cues amidst dynamic environments. Incorporating audio feedback with Facial Expression Recognition (FER) systems via pre-trained or untrained Multimodal Large Language Models (MMLMs) presents 3 significant challenges [20]. The complexities arise from the potential mismatch between facial expressions and audio cues, requiring sophisticated algorithms to accurately synchronize and interpret these multimodal inputs for effective communication.For instance, imagine a scenario where a Facial Expression Recognition (FER) system, trained on standard datasets, encounters a person expressing happiness through a smile while delivering feedback in a cheerful tone. However, due to cultural differences or individual variations, the system misinterprets the expression as sarcasm or discomfort, leading to an incongruent audio response. This highlights the real-life challenge of aligning facial expressions with audio feedback accurately, especially in diverse cultural contexts or for individuals with unique emotional expressions. The creation of a real-time

Facial Expression Recognition (FER) system with Bengali audio feedback represents a crucial advancement in technology, bridging a notable gap in current capabilities and providing a more inclusive solution for diverse user populations. Through the optimization of FER algorithms and the seamless integration of language-specific audio feedback, this research endeavors to enhance communication efficacy and user satisfaction across a wide range of contexts. The practical implications of this thesis extend to vital sectors like customer service, education, and healthcare, where effective communication serves as the cornerstone of success and meaningful interaction.

1.3 Research Objectives

Imagine seamless communication where facial expressions translate into clear Bengali audio descriptions. The objective of this study is to develop a comprehensive emotion recognition system that integrates multimodal data sources, including physiological data (such as heart rate), physical expressions (facial expressions and body language), and linguistic cues (audio and text). Specifically, the research aims to:
1. Combine Multimodal Data: Integrate physiological signals, facial expressions, body language, and speech features to capture both categorical (happy, sad, etc.) and dimensional (valence, arousal) aspects of emotions.
2. Utilize Deep Learning Models: Employ advanced deep learning models, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and hybrid architectures, to improve feature learning and enhance the accuracy of emotion recognition.
3. Implement and Evaluate Frameworks: Develop and evaluate frameworks like EALD-MLLM for analyzing long-sequential and de-identified videos, focusing on non-facial body language cues and their impact on emotion recognition.
4. Leverage Pre-trained Models and Advanced Techniques: enhance emotion recognition through pre-trained models (Wav2Vec2.0 for speech, BERT for text) and advanced FER techniques. Deep learning architectures combined with facial landmark detection and Action Unit (AU) analysis will target high accuracy and capture temporal dynamics of emotions.
5. Compare FER Toolboxes: Analyze and compare the effectiveness of various FER toolboxes, such as Open Face 2.0, Py-Feat, Affdex 2.0, LibreFace, and PyAFAR, in different domains.
6. Benchmark Against State-of-the-Art Models: Compare the performance of multimodal approaches with state-of-the-art single-modal models, including generative models like GPT-3.5 and fine-tuned models like RoBERTa, particularly in the context of translated non-English text data.

By achieving these objectives, the study aims to advance the field of affective computing, providing a robust and accurate emotion recognition system that leverages the strengths of multimodal data integration and cutting-edge deep learning techniques. The integration of physiological signals, facial expressions, body language, and speech features ensures a comprehensive capture of emotional states. Employing advanced deep learning models, particularly CNNs, RNNs, and hybrid architectures, enhances feature learning and recognition accuracy. Developing and evaluating frameworks like EALD-MLLM [22] allows for detailed analysis of long sequential data and non-facial body language cues. Leveraging pre-trained models

like Wav2Vec2.0 and BERT [12] combined with advanced FER techniques, further refines emotion recognition by focusing on speech and text nuances and facial landmarks. Comparing the effectiveness of various FER toolboxes and benchmarking against state-of-the-art models ensures the robustness and versatility of the proposed system, ultimately contributing significantly to multimodal deep learning techniques and improving communication experiences, especially in Bengali-speaking regions.

Chapter 2

Literature Review

The rapid advancements in artificial intelligence (AI) and machine learning (ML) have led to significant progress in various fields, including emotion recognition systems. Emotion recognition, particularly real-time facial expression recognition (FER), has the potential to revolutionize human-computer interaction by enabling machines to understand and respond to human emotions effectively. These systems find applications in numerous domains such as healthcare, education, entertainment, security, and customer service, where understanding human emotions can greatly enhance the user experience and operational efficiency.Real-time FER systems primarily focus on analyzing facial expressions to detect emotions. However, relying solely on facial expressions can be limiting due to various factors such as occlusions, lighting conditions, and individual differences in expressing emotions.Our proposed method combines facial recognition with Bengali voice analysis to bridge communication gaps and enhance real-time emotional understanding.Our literature review explores existing facial recognition systems and how they can be improved by incorporating voice data, particularly in Bengali-speaking regions like Bangladesh and West Bengal. This combined approach can lead to a more nuanced understanding of emotions by analyzing both verbal and non-verbal cues.We further delve into advanced models and deep learning techniques for capturing complex emotional signals from different data sources. The review acknowledges the challenges of privacy, cul- tural variations, and real-time application limitations, while emphasizing the need for versatile datasets and further research in these areas. This research aims to contribute to the development of robust and adaptable emotion recognition systems across diverse linguistic and cultural contexts. This section explores the challenges of recognizing emotions from facial expressions in natural environments, as highlighted in a recent study by Bian et al.(2024) The authors propose that Multimodal Large Language Models (MLLMs) hold promise for improved emotion recognition, particularly in complex situations, by leveraging contextual information to overcome limitations of facial expressions alone . Bian .(2024) also compare characteristics, architectures, and performance of existing FER toolboxes like Open Face 2.0, PyFeat, Affdex 2.0, LibreFace, and PyAFAR [27].On the top, The DISFA dataset, a benchmark for evaluating AU detection in FER tool- boxes, which includes video-induced spontaneous expressions [27].The future of FER lies in developing models that can effectively analyze naturalistic facial expressions and incorporate context. Additionally, building robust systems for facial expres- sion recognition in real-world settings is crucial. Author mentioned some factors 6 such as uncontrolled settings

and lack of background information that can hinder accurate emotion recognition [27]. MLLMs offer potential for FER, but limitations exist. MLLMs may not be able to capture the precise details of facial expressions as well as traditional FER tools. Distinguishing subtle differences in facial expressions, like variations in smiles, could also be challenging for MLLMs. Furthermore, it is uncertain how well MLLMs' understanding of emotional context aligns with human perception. Contextual biases and individual variations in emotional response could affect MLLM accuracy. Multimodal emotion recognition uses different data sources, such as facial expressions, speech signals, and text data, to improve the accuracy and reliability of emotion recognition systems. A recent study by Freshworks (2022) that proposes a novel approach for multimodal emotion recognition using large pre-trained models with cross-modal attention [20]. It underscores the significance of self-supervised pre-training in machine learning, particularly in speech and NLP domains, where deep learning models have surpassed classical methods. Leveraging models like Wav2Vec2.0 for speech and BERT for text [20], trained on vast unlabeled data, the approach fine-tunes these models with labeled data to enhance classification accuracy [20]. The architecture integrates audio and text encoders with cross-modal attention mechanisms to align features, extracting interactive information crucial for emotion recognition. Combining these models with a multifaceted attention mechanism improves the system's ability to learn the interaction of speech and textual information, leading to more accurate emotion recognition. This approach showed a significant 1.88 percent accuracy improvement over previous state-of-the-art methods on the IEMOCAP dataset [20].Limiting the analysis to audio and text data, the study could overlook the potential benefits of incorporating facial expressions for a more comprehensive picture of emotional states. Accurate emotion recognition hinges on capturing both the spatial configuration of facial features (e.g., eyebrow position, mouth curvature) and the temporal dy- namics of expressions (changes over time). A recent survey by Monica and Mary (2022) examining cutting-edge methods in Facial Emotion Recognition (FER) using deep learning architectures [24] underscores this very point. Their work emphasizes the importance of capturing both the static and dynamic aspects of facial expres- sions for accurate emotion recognition. The survey delves into specific techniques to address these spatial and temporal dynamics, including Inception-ResNet for 3D feature extraction from videos, Long Short-Term Memory (LSTM) networks for capturing temporal information, and CNN-RNN hybrids that combine these ap- proaches [24]. The paper highlights the use of facial landmark detection to pinpoint key regions on the face, further improving recognition accuracy. By identifying these landmarks, the model can focus on crucial emotional cues and ensure training on precise data.The study acknowledges limitations in the proposed method, including generalizability to new data, handling large datasets and real-time video processing (scalability concerns), and high computational cost due to complex 3D-CNNs and LSTMs [24]. Additionally, the method's accuracy is sensitive to precise facial land- mark localization, which can be problematic in scenarios with challenging detection. Li, Deng et al. (2024) introduces EALD, a groundbreaking dataset designed for emo- tion analysis in long videos while prioritizing user privacy through de-identification [28]. The EALD dataset features anonymized athlete interviews (averaging 7 min- 7 utes each) and incorporates annotations of non-facial body language cues to provide a richer understanding of emotions . An MLLM framework within EALD-MLLM outperforms

traditional single-modal models, even when analyzing emotions for the first time (zero-shot learning) [28]. MLLMs excel at comprehending and analyzing emotions in new scenarios without requiring retraining. It highlights the importance of non-facial body language (NFBL) as a privacy-preserving indicator of emotional states. Integrating biosignals(e.g.,electroencephalogram (EEG), heart rate, skin conductance),[23]with facial expressions and audio data could lead to a more nuanced under- standing of true emotions. A recent review by Maithri et al. (2022) that examines the state-of-the-art in Automated Emotion Recognition (ER) systems , A key fo- cus is multimodal fusion approaches, which combine multiple modalities like EEG with facial expressions to improve the accuracy and robustness of ER systems [23].It highlights the application of both machine learning (ML) and deep learning (DL) techniques in analyzing these modalities.Real-time ER with multimodal systems re- mains a challenge due to processing requirements for multiple data streams .Hence this study offeres a more holistic understanding of emotions by integrating physio- logical, visual, and auditory data . Accurately detecting emotions using large language models (LLMs) presents a chal- lenge, particularly when considering the link to video content and the regional con- text within the Bangla language. A research by Venkatakrishnan et al. (2023) that investigates the emotion detection capabilities of large language models (LLMs) in the context of the Middle East and deeply focuses on two transformer-based models: GPT-3.5 and RoBERTa [26]. A considering side for this research indicates on struggle with fine-grained classification due to its generative nature, potential language and translation issues, high resource requirements, and the need for further research on bias and cultural sensitivity .It finds GPT-3.5 excels at generating relevant text but struggles with fine-grained emotion classification, while RoBERTa is more accurate for specific emotions .This study analyzes emotional responses (Persian & Turkish) to the murder of Zhina (Mahsa) Amini and the earthquake in Turkey/Syria using a 10,000-tweet dataset (5k/event) collected from 6 million via SNScrapper [26]. This review by Wang et al. (2022) highlights a critical shift in emotion recognition: the establishment of standardized emotion categorization methods and evaluation metrics [25]. These techniques move the field away from subjective and inconsis- tent approaches, paving the way for a more scientific and reliable way to recognize emotions [25].Furthermore, the research explores the potential of both single-source (unimodal) and combined-source (multimodal) approaches for capturing emotions in real-world scenarios .This opens exciting possibilities for analyzing complex emotions within the natural context of daily life, going beyond the limitations of traditional methods. Emotional recognition is on the rise, supported by artificial intelligence and machine learning. While the look is promising, their limitations are clear. The future lies in combining them with other data sources, such as voice and text messages, to get a complete picture. Research suggests using large language models to understand context, but there are still challenges in capturing details and aligning machine understanding with human perception. Advanced deep learning architectures can capture both static and dynamic aspects of expressions, while the integration of biosignals such as EEG can provide a more nuanced view of emotions. Overcoming challenges such as scalability and privacy issues is critical. New research using datasets similar to EALD shows the potential of non-facial body language cues. This, along with advances in multimodal fusion and deep learning, paves the way for older and more adaptive emotion recognition systems that can understand emotions in different contexts.

The key to the future is leveraging multiple data sources, incorporating context and using advanced deep learning techniques. By addressing existing barriers and ethical considerations, we can develop systems that provide a more accurate and nuanced understanding of human emotions..

2.1 Real-Time Facial Expression Recognition(FER)

A Real time facial expression recognition (FER) system refers to systems designed to analyze facial expressions in uncontrolled, real-world settings. These systems aim to capture the complexity and spontaneity of emotions that occur in daily life, such as those seen in videos, vlogs, or documentaries. FER systems are particularly challenging because they must account for variables like changes in lighting, head orientation, background complexity, and facial occlusions.

2.2 FER Architecture

Bian et al. (2024) propose a sophisticated architecture for Naturalistic Facial Expression Recognition (FER) systems, comprising five hierarchical layers: the Face Detection Layer, Feature Extraction Layer, Emotion Prediction Layer, Contextual Processing Layer, and Multimodal Integration Layer [27]. The Face Detection Layer identifies and tracks faces in unconstrained environments, effectively handling challenges such as variable lighting and occlusions. The Feature Extraction Layer utilizes Convolutional Neural Networks (CNNs) to extract critical facial features and action units from the detected faces. Following this, the Emotion Prediction Layer classifies these features into specific emotional categories using pre-trained models. The Contextual Processing Layer enhances adaptability by incorporating external data, including body language and spoken words, to refine emotion inferences. Finally, the Multimodal Integration Layer synthesizes outputs from diverse sources, such as audio and text, employing Multimodal Large Language Models (MLLMs) to achieve a comprehensive emotional understanding. Additionally, Management Capabilities and Security Capabilities are integrated across all layers to address real-time performance, privacy, and security issues, ensuring that the FER system operates efficiently in naturalistic settings.

2.3 Related Works

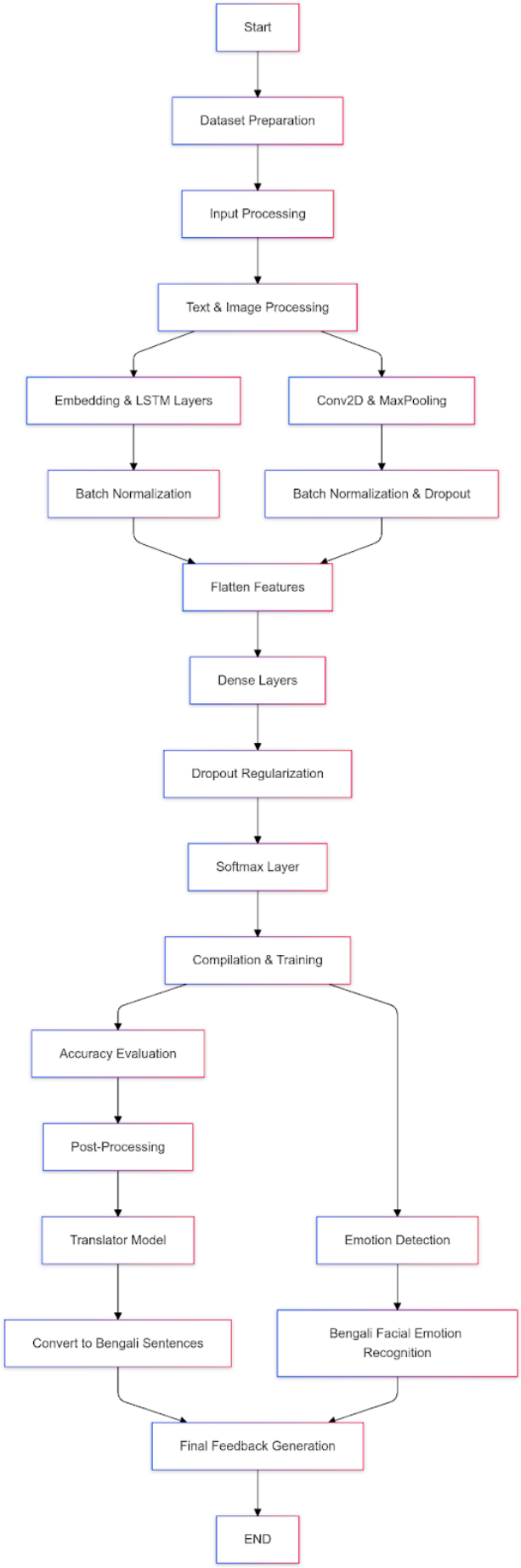
Real-time FER systems often rely solely on facial expressions, which can be limited. We propose combining facial recognition with Bengali voice analysis for better emotional understanding. This part reviews previous work on real-time FER systems.
Bian (2024) conducted a comprehensive review of existing facial expression recognition (FER) systems and explored the limitations of relying solely on facial expressions for emotion recognition. The research acknowledged the challenges of using MLLMs, such as their difficulty in capturing precise details in facial expressions and their sensitivity to contextual biases. The architecture of Multimodal Large Language Models (MM-LLMs) typically comprises five key components: the Modality Encoder, Input Projector, LLM Backbone, Output Projector, and Modality Generator. The Modality Encoder processes inputs from various modalities, such as images,

audio, and text, transforming them into feature representations. These features are then aligned with the text space by the Input Projector, enabling the LLM Backbone to perform reasoning and decision-making tasks. The Output Projector subsequently maps the processed features back to the appropriate output format, which is finally generated by the Modality Generator in the desired modality (e.g., images, audio). This architecture allows efficient training by freezing most of the larger components and focusing optimization on a lighter projection layer.The study also discussed the potential limitations in aligning machine-generated emotional understanding with human perception. Large pre-trained models with cross-modal attention mechanisms and self-supervised pre-training in speech and natural language processing (NLP) are particularly promising in this context [20]. The multimodal emotion recognition model proposed by Freshworks (2022) leverages large pre-trained models and cross-modal attention mechanisms to enhance emotion recognition accuracy. The architecture includes an audio encoder base, using the Wav2Vec2.0 model pre-trained on 53,000 hours of audio to extract robust speech features. This is followed by an audio feature encoder, which employs two 1D convolution layers and a bidirectional LSTM (BLSTM) to reduce the frame rate and capture temporal dependencies. Similarly, a text encoder base utilizes the BERT model to generate rich contextual embeddings from text data. These embeddings are projected to a lower dimension using a convolutional layer for compatibility with the audio features.The model's core lies in its cross-modal attention mechanism, consisting of two modules: CMA-1, which aligns audio with text, and CMA-2, which aligns text with audio. These modules use multi-head attention to capture bidirectional interactions between the two modalities, facilitating better feature alignment. The output features from both cross-modal attention modules are processed through a statistics pooling layer, which computes the mean and standard deviation to generate utterance-level feature representations[20]. These pooled audio and text features are concatenated and passed to a classification layer to predict the emotion label. Both aim to achieve better emotion recognition through multimodal data but differ in their data integration focus.Bian (2024) and Freshworks (2022) both enhance emotion recognition using multimodal approaches but with distinct focuses. Bian (2024) highlights the limitations of relying on facial expressions alone, proposing Multimodal Large Language Models (MM-LLMs) to integrate images, audio, and text for comprehensive emotional understanding[27]. Freshworks (2022), on the other hand, centers on speech-text integration using cross-modal attention to align features between these two modalities.

Chapter 3

Methodology

The research employs a dual-stream deep mechanism, a mixture of face and text processing to generate speech feedback in terms of face emotion. Face and Bengali text datasets are initially prepared and processed. Text processing employs embedding layers and LSTM networks, and face processing employs feature extraction through convolutional layers. Two sets of features then merge and go through dense layers, dropout regularization, and a softmax classifier for emotion classification. Model training and accuracy testing occur, and detected emotion is translated into a sequence of Bengali sentences. The system then generates speech feedback, offering real-time, emotion-sensitive feedback.


Start
Dataset Preparation
Input Processing
Text & Image Processing
Embedding & LSTM Layers
Conv2D & MaxPooling
Batch Normalization
Batch Normalization & Dropout
Flatten Features
Dense Layers
Dropout Regularization
Softmax Layer
Compilation & Training
Accuracy Evaluation
Post-Processing
Translator Model
Emotion Detection
Convert to Bengali Sentences
Bengali Facial Emotion Recognition
Final Feedback Generation
END

3.1 Input data

For this research on **Real Time Facial Emotion Recognition with Bangla Audio Feed**, we have utilized two main datasets: an image dataset containing facial expressions and an audio dataset with emotional audio features.

Image Dataset

The image dataset consists of labeled facial expression images, each annotated with the corresponding emotion label, gender, and age range of the individual depicted. The key attributes of this dataset include:

Image Name: Each image file is identified by a unique filename .

Expression Label: This attribute represents the emotional expression portrayed in the image. Categories include emotions like 'angry', 'happy', and 'sad'.

Emotion Label: Each expression is further encoded into a numerical label for machine learning purposes (e.g., 0 for 'angry').

Demographic Information: In addition to the facial expression, each image is annotated with the gender (e.g., 'Male' or 'Female') and an age range (e.g., '(25-32)') of the individual.

	sl	img_name	expression	label	gender	age
0	1	11024.jpg	angry	0	Female	(25-32)
1	2	13395.jpg	angry	0	Female	(38-43)
2	3	15194.jpg	angry	0	Male	(25-32)
3	4	14945.jpg	angry	0	Male	(4-6)
4	5	11158.jpg	angry	0	Female	(25-32)

Figure 3.1: Demographic Information sample from the dataset.

Each audio sample is associated with a corresponding emotion label, which aligns with the emotion in the facial expression dataset.This sample collected from the West Bengal and Bangladesh content over the internet .

	Unnamed: 0	path	label
0	1	Surprise/1bd930d6a1c717c11be33db74823f661cb53f...	Surprise
1	2	Surprise/cropped_emotions.100096~12fffff.png	Surprise
2	3	Surprise/0df0e470e33093f5b72a8197fa209d684032c...	Surprise
3	4	Surprise/cropped_emotions.260779~12fffff.png	Surprise
4	5	Surprise/cropped_emotions.263616~12fffff.png	Surprise

Figure 3.2: Label Path sample from the dataset.

This multimodal dataset allows for the exploration of correlations between visual and audio cues in emotion recognition tasks.

3.2 Data pre-processing

In this study, we compiled a comprehensive dataset consisting of 62,058 facial expression images to develop a Facial Expression Recognition (FER) system for emotion classification. The dataset was created by combining two well-known FER datasets:
FER 2013 – This dataset contains 35,887 grayscale images, each with a fixed resolution of 48×48 pixels. It is a widely used benchmark dataset for facial expression recognition, covering seven basic emotions: Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise.
AffectNet V2 – This dataset originally consists of 96×96 pixel color images, with a total of 26,171 samples selected for this study. AffectNet V2 is one of the largest facial expression datasets, capturing a diverse range of real-world expressions.

3.2.1 Preprocessing and Image Resizing:

Since the FER 2013 dataset has images of 48×48 pixels, while AffectNet V2 images are of 96×96 pixels, we standardized the input size by resizing all AffectNet V2 images to 48x48 pixels. This ensures consistency across the dataset and allows both datasets to be effectively used together for training the FER system.

Dataset	**Original Image Size**	**Number of Images**	**Preprocessing Steps**
FER 2013	48×48 (grayscale)	35,887	No resizing needed
AffectNet V2	96×96 (color)	26,171	Resized to 48×48

Table 3.1: Dataset Overview

By combining these datasets, we established a more comprehensive and varied dataset, guaranteeing an equitable distribution of face expressions from several sources. This integration improves the model's capacity to generalise well to unfamiliar face expressions in practical situations.

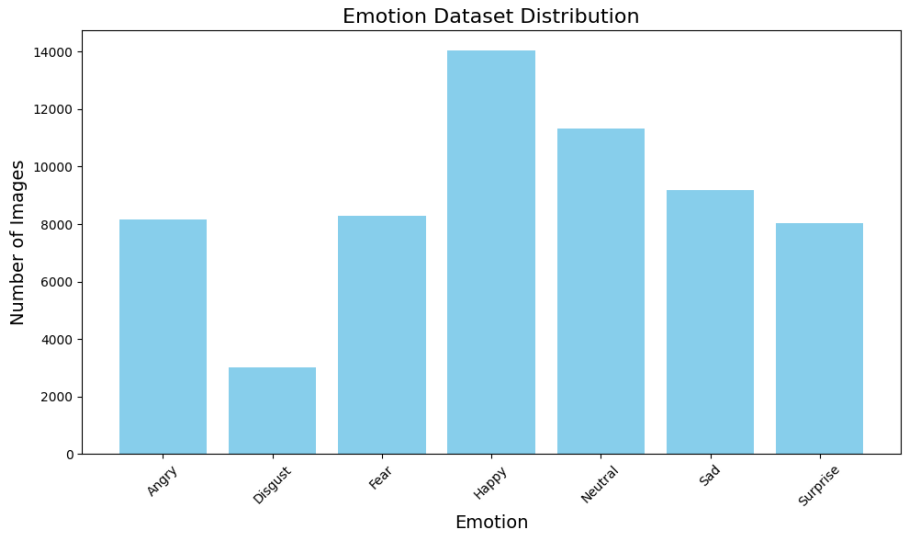


Figure 3.3: Emotion Distribution in Dataset.

```
    print(f"Number of emotions: {num_emotions}")
    print("Emotion names:", unique_emotions)

✓ 2.2s

Number of emotions: 7
Emotion names: ['angry' 'disgust' 'fear' 'happy' 'neutral' 'sad' 'surprise']
```

Figure 3.4: Number of emotions.

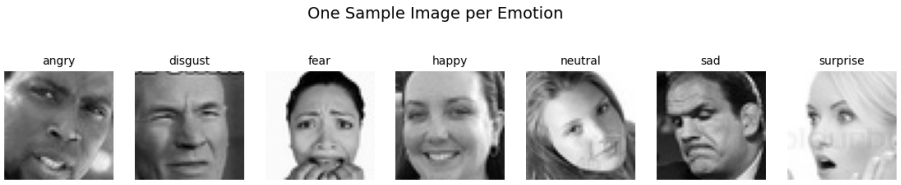


Figure 3.5: One Sample Image per Emotion.

A notable disparity in the quantity of photos each class was identified, especially within the 'disgust' class, which had the least examples.

Addressing Class Imbalance: Class imbalance is a typical problem in machine learning, when one class has more samples than another. When the dataset is skewed, the model may learn to predict the majority class more often, producing biased predictions and poor generalisation. In the case of FER, under-representation of some emotions may have an influence on model accuracy, particularly for emotions such as 'disgust', which are critical to the system's performance. To guarantee that the model does neither overfit or underfit owing to this imbalance, we used class weighting and balancing approaches.

Default Class Weights: At first, each class had the same weight of 1.0, which meant that no class was more important than the others. Most of the time, these default weights are used when the dataset isn't significantly imbalanced. However,

they might not be the best choice when some feelings aren't well represented. It's possible that this idea of equal weight made the model do badly on the mood classes that happened less often.

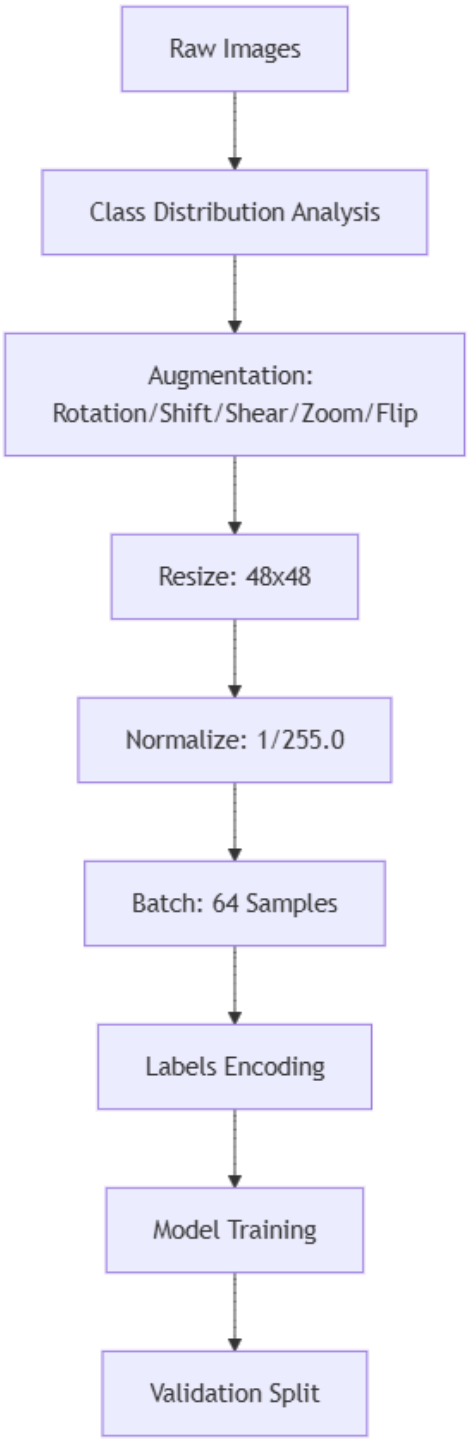
In order to rectify the imbalance, we determined the class weights by analysing the frequency of each class in the dataset. Class weighting is designed to elevate the significance of minority classes, such as "disgust," to ensure that the model places a greater emphasis on them during the training process. The class weights that were calculated are as follows:

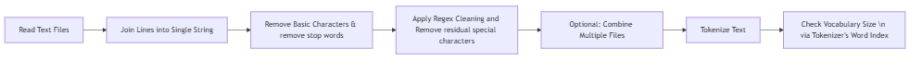
Class	**Value**
Anger	1.0897
Disgust	2.9481
Fear	1.0766
Happy	0.6311
Neutral	0.7865
Sad	0.9599
Surprise	1.0910

Table 3.2: Class Values

The class 'disgust' has the highest weight (2.9481) due to its under-representation in the dataset, as evidenced by these computed weights. In contrast, the class 'glad' has the lowest weight (0.6311) due to the fact that it contains the most samples in the dataset. These weights have been adjusted to prevent the model from favouring the more abundant classes disproportionately and to improve its generalisation across all emotions, particularly those that are under-represented.
Balancing the dataset, together with estimated class weights, helps to reduce overfitting and underfitting. Overfitting happens when a model performs well on training data but fails to generalise to new data, which is generally caused by the model memorising patterns unique to the majority class. Underfitting occurs when a model fails to capture the underlying patterns of minority classes, such as 'disgust'. Using class weights forces the model to treat minority classes more equitably, resulting in greater generalisation across all emotional categories. In this dataset, class balancing using class weights is especially important for enhancing the model's performance on the 'disgust' class. This guarantees that the model learns enough characteristics for this emotion, lowering the likelihood of incorrect categorisation or misclassification during testing and real-world application.
Also,Each image undergoes a systematic preprocessing pipeline before model training. First, images are loaded from class-specific directories in train and validation splits, and class distributions are analyzed to identify potential imbalances, a practice critical for mitigating bias [16]. To enhance model robustness, images are augmented in real-time using geometric transformations, including random rotations ($\pm 20°$), horizontal/vertical shifts (20 percent of image width/height), shear (20 percent), and zoom (20 percent), which simulate diverse viewing conditions and reduce overfitting [25]. Additionally, horizontal flipping is applied to improve invariance with object orientation. All images are resized to a uniform resolution of 48×48 pixels to standardize input dimensions while preserving discriminative features . Pixel values are normalized to the range [0, 1] by rescaling (dividing by 255), a stan-

dard practice for stabilizing neural network training [13]. Labels are converted to one-hot encoded vectors for multi-class classification. Finally, images are batched into groups of 32 to optimize memory usage and gradient computation during training [18]. This preprocessing ensures the model generalizes effectively to unseen data while retaining semantic relevance in augmented samples[19]).The dataset for feedback generation was preprocessed by consolidating raw text files into a single string, followed by systematic removal of unwanted characters, including newline (/n), carriage return (/r), Unicode formatting marks (e.g., /ufeff), and typographical quotation marks (“, ”, ‘, ’) using string replacement and regular expressions [7]. Residual special characters were eliminated through regex-based sanitization to ensure textual uniformity, a common practice in NLP pipelines. [1]. Tokenization was performed using a standard tokenizer to split the text into discrete linguistic units, with vocabulary size derived from the tokenizer's word index [23]. This preprocessing workflow aligns with established methodologies for preparing unstructured text data for computational analysis .[17]


Raw Images
Class Distribution Analysis
Augmentation:
Rotation/Shift/Shear/Zoom/Flip
Resize: 48x48
Normalize: 1/255.0
Batch: 64 Samples
Labels Encoding
Model Training
Validation Split



3.3 Multimodal Dataset

The multi-modal dataset integrates two distinct data modalities: facial expression images and Bengali textual feedback, enabling joint learning of visual emotion recognition and contextually relevant linguistic responses. This fusion aligns with frameworks for multi-modal learning that leverage complementary information across modalities [15].

Data Synchronization and Alignment:

1. Cross-modal label mapping: Emotion labels from the image dataset (e.g., angry-0, happy-2) were mapped to semantically equivalent Bengali words (e.g., রাগান্বিত and খুশি) in the text corpus.This ensures alignment between visual predictions and linguistic outputs, a critical step for coherent multimodal systems [9].

2. For real-time applications, timestamps were embedded to synchronize audio-text feedback generation with detected facial expressions, ensuring temporal coherence [21].

Preprocessing Pipeline:

1. Image Modality:

   - Geometric Standardization: Images resized to 48×48 pixels [10] and normalized to [0, 1] [13].
   - Augmentation: Applied rotations, shifts, and flips to simulate real-world variability [14].

2. Text Modality: :

   - Sanitization: Removed Unicode artifacts and typographical noise via regex .[7]
   - Tokenization: Split text into subword units using a vocabulary of 7k+ tokens, optimized for LSTM training [6].

3. Joint Representation:

    - Feature Concatenation: Visual features (CNN embeddings) and textual features (LSTM hidden states) were concatenated to form a unified input vector for downstream tasks.[11].
    - Batch Alignment: Batches were structured to include paired image-text samples, ensuring synchronized updates during training [15].

Chapter 4

Implementation

In this section, we detail the implementation of our multimodal approach for predicting facial emotions with Bengali Audio feedback.
The model for emotion detection utilizes a face emotion analysis via a convolutional neural network (CNN) model. Grayscale face images of 48x48 pixels are processed in its first layer. In feature extraction, a sequence of a few layers of ReLU activation, same-padding, and operations of max-pooling is performed for the extraction of spatial information. For enhancing its stability and for overfitting compensation, operations of batch normalization and dropout are added in between. Features extracted in such a manner go through flattening and then pass through its head's fully connected dense layers for classification. For output, a softmax function is added in its output layer for emotion classification. For efficient training and multi-class classification, a model is compiled with an Adam optimizer and categorical cross-entropy loss.
The model for producing feedback employs a stacked LSTM model for sequence processing of text. It processes an 151-sequence processing layer, projecting each token onto a 100-dimensional dense vector via an embedding layer. There are three LSTM processing sequential relations with 128 units, with two producing full sequences and one producing only its final state. There is training stability via batch normalization. There is an output layer with 7173 softmax-activated units producing distributions over target categories. Model training employs the Adam optimizer and categorical cross-entropy loss. After training, post-processing predicted sequences of tokens use a translator model to produce coherent Bengali feedback, producing readable and natural output.

4.1 FER Model

4.1.1 Input Layer:

The model accepts input images with a shape of (48, 48, 1), representing 48x48 pixels with a single color channel (grayscale). This layer defines the input dimensions for the image data that will be processed by the CNN.

4.1.2 Extractor (CNN Part):

i) Convolutional Layers (Conv2D): The model uses multiple convolutional layers with varying numbers of filters. Each convolutional layer applies filters to extract features from the input image. The filters' size and number can be defined through the kernel sizes and filters parameters, respectively. The convolutional layers apply the specified activation function (default:'relu') for nonlinear transformations.

ii) Padding: The convolutional layers use'same' padding to ensure that the output size is the same as the input size, with the necessary padding applied to the borders.

iii) Strides and Dilation: Strides control the step size of the filter while scanning the input image, and dilation (currently applied in the first layer) allows the filter to cover a larger receptive field without increasing the number of parameters.

iv) MaxPooling Layers (MaxPooling2D): The model applies pooling operations to downsample the feature maps and reduce their spatial dimensions, helping reduce computation and improve generalization.

v) Batch Normalization: After each convolution, batch normalization is applied to stabilize and speed up the training by normalizing the layer's output.

vi) Dropout Layers: Dropout is introduced between convolutional layers if the dropout_rate is specified, which helps in regularization and preventing overfitting during training.

4.1.3 Flattening:

After the feature extraction phase, the multi-dimensional feature maps are flattened into a one-dimensional vector, preparing the extracted features for input to the fully connected layers in the classification part.

4.1.4 Classification Head (ANN Part):

i) Dense Layers: The flattened features are passed through a series of fully connected (dense) layers. The number of units in these layers is specified by the dense units parameter. Each dense layer uses the specified activation function (default:'relu').

ii) Dropout Layers: Dropout is optionally added to the dense layers if a dropout rate is defined, helping to regularize the model further.

4.1.5 Output Layer:

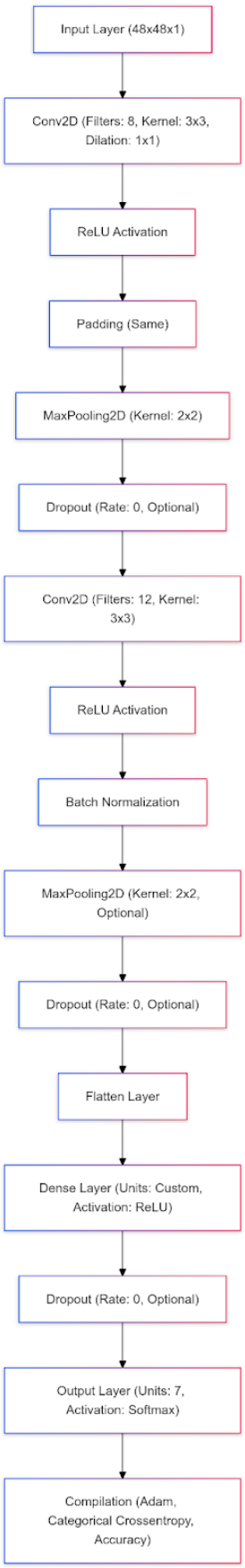
The final dense layer has output classes units corresponding to the number of output classes, with a'softmax' activation function to output the probability distribution across the predicted classes.

4.1.6 Compilation:

The model is compiled using the Adam optimizer, which is known for its adaptive learning rate. The categorical cross-entropy loss function is chosen because it is appropriate for multi-class classification problems. The metric used to evaluate the model's performance is accuracy.

4.1.7 Model Overview:

The architecture of this model is built around a CNN feature extractor followed by a classification head composed of dense layers. The convolutional layers learn spatial features of the input images, such as edges, textures, and shapes, which are then used by the fully connected layers to classify the image into one of the output classes categories.
Dropout and batch normalization are incorporated throughout the architecture to regularize the model and stabilize training, preventing overfitting and ensuring faster convergence.


Input Layer (48x48x1)
Conv2D (Filters: 8, Kernel: 3x3, Dilation: 1x1)
ReLU Activation
Padding (Same)
MaxPooling2D (Kernel: 2x2)
Dropout (Rate: 0, Optional)
Conv2D (Filters: 12, Kernel: 3x3)
ReLU Activation
Batch Normalization
MaxPooling2D (Kernel: 2x2, Optional)
Dropout (Rate: 0, Optional)
Flatten Layer
Dense Layer (Units: Custom, Activation: ReLU)
Dropout (Rate: 0, Optional)
Output Layer (Units: 7, Activation: Softmax)
Compilation (Adam, Categorical Crossentropy, Accuracy)

4.2 Feedback Generation:

4.2.1 Model Architecture:

- **Input Layer:**

  The model accepts input sequences of shape (151,), where each element in the sequence represents a token. These tokens are fed into an Embedding layer, which maps the input tokens to dense vectors of size 100. The `input_dim=14268` indicates the vocabulary size, representing the total unique tokens in the input data.

- **Embedding Layer:**

  The Embedding layer transforms sparse input tokens into dense 100-dimensional vector representations. This allows the model to learn meaningful representations of the input tokens during training.

- **Recurrent Layers:**

  Three stacked LSTM (Long Short-Term Memory) layers, each with 128 units, are used to capture sequential dependencies and temporal relationships in the input data. The first two LSTM layers have `return_sequences=True`, which ensures that they output the entire sequence for subsequent layers to process. The third LSTM layer does not return sequences, outputting only the final hidden state to pass to the next layer. These layers are essential for understanding the context in sequential data and are particularly effective for tasks involving language and time-series data.

- **Batch Normalization:**

  A BatchNormalization layer is added after the LSTMs to normalize the outputs. This stabilizes and accelerates the training process by reducing internal covariate shift.

- **Output Layer:**

  A Dense layer with 7173 units and a softmax activation function is the final layer. The 7173 units correspond to the number of target classes, each representing a unique output token. The softmax function generates a probability distribution over all classes, enabling the model to predict the most likely output token.

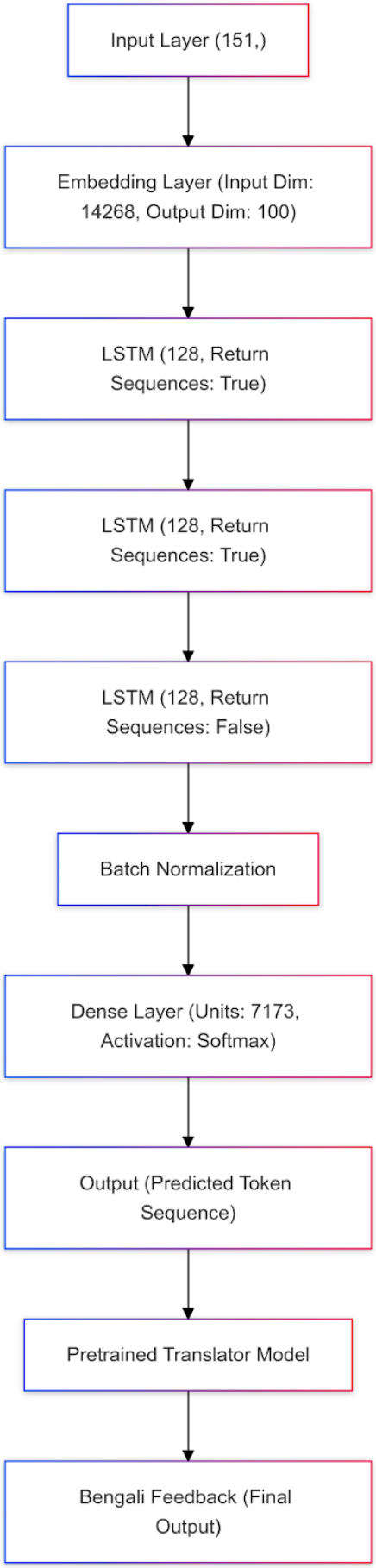
4.2.2 Compilation and Training:

The model is compiled with a suitable optimizer (e.g., Adam) and a loss function such as categorical crossentropy.During training, the model learns to map input token sequences to corresponding output token probabilities effectively.

4.2.3 Post-Processing:

Once the model is trained, the outputs are passed through a pretrained translator model to generate Bengali feedback.The translator model takes the predicted

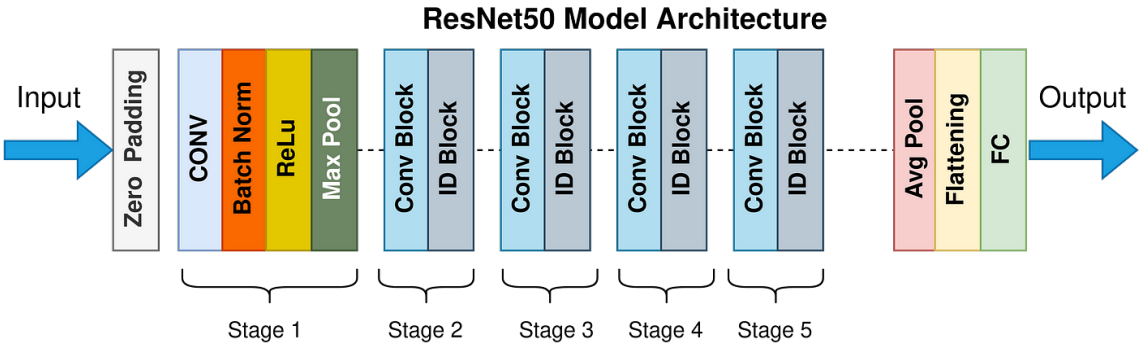
sequence of tokens and converts them into coherent Bengali sentences.This post-processing step is crucial for ensuring the output is natural and interpretable in the Bengali language.


Input Layer (151,)
Embedding Layer (Input Dim: 14268, Output Dim: 100)
LSTM (128, Return Sequences: True)
LSTM (128, Return Sequences: True)
LSTM (128, Return Sequences: False)
Batch Normalization
Dense Layer (Units: 7173, Activation: Softmax)
Output (Predicted Token Sequence)
Pretrained Translator Model
Bengali Feedback (Final Output)

4.3 Pretrained Model:

4.3.1 ResNet 50

ResNet-50, developed by Microsoft Research, is a powerful deep convolutional neural network designed for image classification. It is built on the "Residual Network" architecture and features 50 layers, making it highly effective in recognizing images by leveraging residual connections. These connections enable the model to map inputs to desired outputs more efficiently, facilitating better learning and improving performance compared to earlier stages.
The architecture of ResNet-50 begins with convolutional layers to extract features from input images. It utilizes two primary types of blocks for processing and transforming data: the identity block and the convolutional block. The final classification is handled by fully connected layers. A key innovation of ResNet-50 is its ability to address the vanishing gradient problem, allowing it to maintain gradient stability even with very deep networks of up to thousands of layers—an achievement beyond the reach of many other models.
ResNet-50 is widely preferred due to its simplified training process, enabled by residual mapping, which manages the complexity of deep learning tasks effectively. Most of its core components remain consistent across various configurations of ResNet. For example, residual connections are introduced after blocks comprising convolutional and max pooling layers. In constructing ResNet-50, input images are typically 224x224 pixels, and the architecture includes optimized hyperparameters, such as learning rates ranging from 0.0001 to 0.1 and dropout rates from 0.0 to 0.9, to improve optimization and reduce overfitting. Testing these configurations over multiple iterations helped identify the best settings for achieving optimal performance. The primary difference among ResNet variants lies in the size of input images and the depth of convolutional layers, which will be explored further in the implementation section.



4.4 Application Best_FER

The integration of real-time face analysis technology in therapeutic and customer service settings has been proven to have significant potential, particularly for communities speaking Bengali. With such technology, therapists can assess the emotional state of patients with post-traumatic stress disorder (PTSD), depression, or autism during virtual consultation sessions. By reading micro-facial cues, the system

creates actionable information for clinicians, allowing for personalized and effective care. This practice conforms with studies emphasizing the importance of non-verbal cues in mental state assessment, particularly for verbally restricted subjects.[2]

In customer service, real-time face expression analysis can be used to monitor representatives' state of emotion during contact with them. By reading for cues of dissatisfaction or empathy, the system can provide real-time feedback in Bengali, allowing for improvement in terms of happiness and communication. This use conforms with studies emphasizing the importance of empathetic communication and emotional intelligence in terms of impact.[4]

For deaf and visually impaired individuals, technology can interpret face expression and provide audio feedback in Bengali, allowing for social contact and increased access. For non-verbal individuals, such a feature is particularly beneficial, allowing for communication gaps and increased access for all. Research emphasizes the importance of such assistive technology in enabling persons with disabilities to integrate into work and social environments.[3]

Additionally, the development of chatbots and virtual assistants in the Bengali language that respond dynamically in relation to moods of users is a significant breakthrough in computer-user interaction. With an ability to scan for facial cues, such a system can make an interaction empathetic and natural. All such breakthroughs have a basis in studies in affective computing, with an emphasis placed in emotion-aware systems for enriching a user's experience.[29]

Lastly, adaptability of the system in reading and providing insights in any language ensures its use anywhere in the world. With such adaptability, its use can be scaled in any language and cultural environment, a fact supported in studies in cross-cultural communications and AI-sensitive cultures' requirements.[8] By overcoming language and cultural barriers, such technology can become universally adopted and have an impact.

4.5 Future work

The proposed facial emotion recognition (FER) system with Bengali audio feedback has demonstrated promising results, achieving high accuracy and robustness in emotion detection. However, there are several areas for future exploration and improvement to enhance the system's performance, applicability, and scalability. One key area is dataset refinement and expansion, where collecting more diverse and culturally representative data, incorporating additional emotion classes, and performing cross-dataset validation can improve generalizability. Another important direction is the integration of psychological signals, such as physiological and behavioral cues, to develop a multimodal emotion recognition system with real-time signal processing. Additionally, exploring advanced architectures, such as Vision Transformers or Graph Neural Networks, and adopting self-supervised learning techniques can optimize model performance. Leveraging pre-trained models through transfer learning and ensembling can further improve accuracy while reducing training time. Advanced preprocessing techniques, including sophisticated data augmentation, facial landmark detection, and noise reduction, can enhance the system's ability to detect

subtle expressions. Implementing a universal quality checker and confidence scoring mechanism can help assess input quality and prediction reliability. Cross-cultural adaptation and multilingual support are also necessary to ensure the system's effectiveness across different populations. Real-world deployment and user feedback will be crucial for identifying practical challenges and refining the system for various applications, while ethical considerations, such as bias mitigation and privacy protection, must be addressed to ensure responsible use. Lastly, integrating the system with virtual assistants and augmented reality applications can expand its usability in interactive environments. By addressing these future directions, the FER system can become more robust, versatile, and ethically sound, providing meaningful and accurate feedback across diverse real-world scenarios.

Chapter 5

Results

5.1 Performance of pre-Trained Model:

5.1.1 Resnet 50

We combined the both FER 2013 and Affectnet v2 dataset.Prior to training, input images were resized to 48x48. The output layer of the ResNet model was modified to accommodate seven classes, corresponding to the severity levels of emotion. The model was trained for 10 epochs, and the resulting training progress is visualized in the figure below.

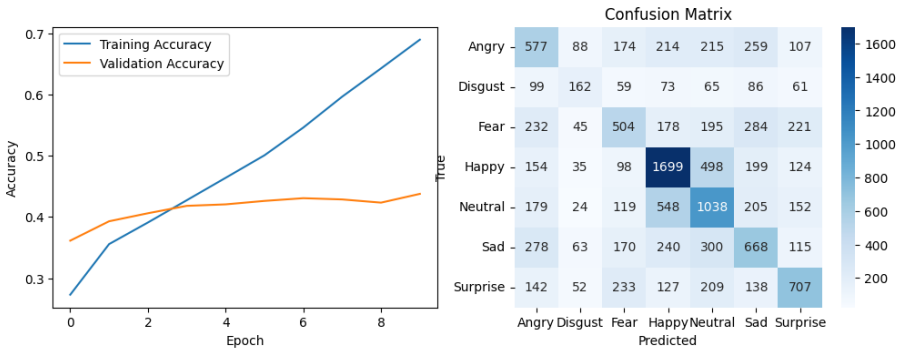


Figure 5.1: ResNet-50 Result

As depicted, the training accuracy (represented by the blue line) steadily increases over the epochs, culminating in a final training accuracy of 68.93%. The validation accuracy (represented by the orange line) also shows improvement, reaching a final validation accuracy of 43.36%. The graph illustrates the general trend of increasing accuracy with training time.

5.2 Performance of Only Fer 2013 Dataset:

Here, we only used the Fer 2013 dataset, which has 35887 image files. All the image sizes were 48*48. The original size of the image was 48*48.The output layer of the EfficientB0 Model was modified to accommodate seven classes, corresponding to

the severity levels of emotion. The model was trained for only 10 epochs because after that the validation accuracy was dropping, and it shows the training accuracy of 83.63% and validation accuracy of 61.38%. The resulting training progress is visualized in the figure below.

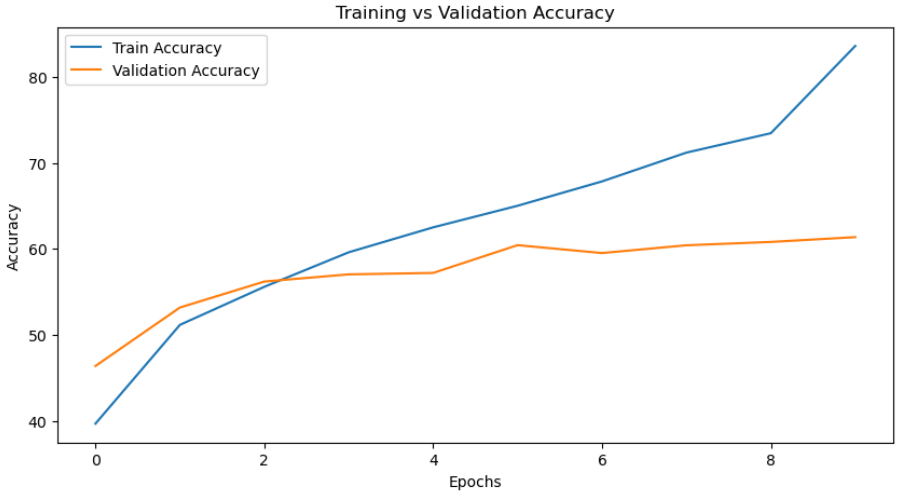


Figure 5.2: Training vs Validation Accuracy

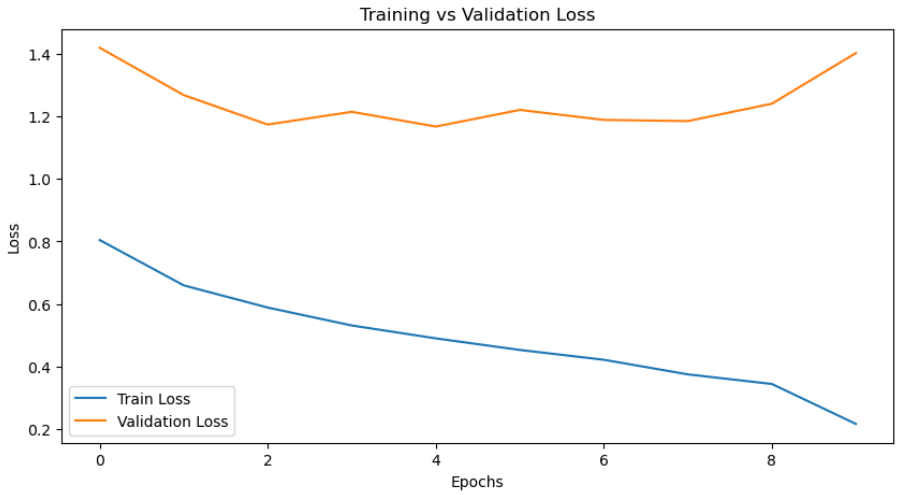


Figure 5.3: Training vs Validation Loss

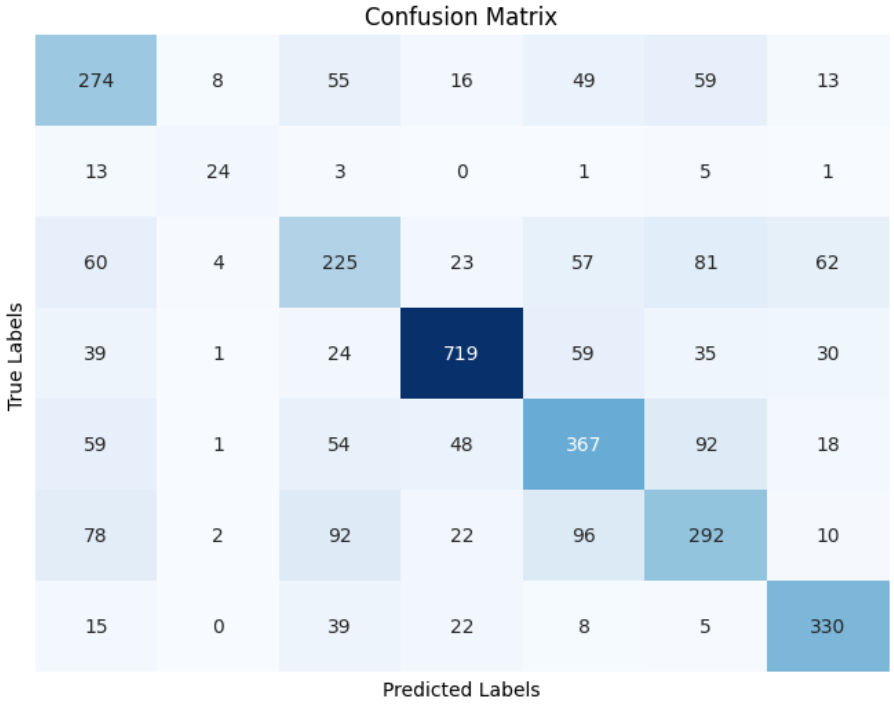


Figure 5.4: Confusion Matrix

Class	**Precision**	**Recall**	**F1-score**	**Support**
Angry	0.51	0.58	0.54	474
Disgust	0.60	0.51	0.55	47
Fear	0.46	0.44	0.45	512
Happy	0.85	0.79	0.82	907
Neutral	0.58	0.57	0.58	639
Sad	0.51	0.49	0.50	592
Surprise	0.71	0.79	0.75	419
Accuracy	0.62			3590
Macro avg	0.60	0.60	0.60	3590
Weighted avg	0.62	0.62	0.62	3590

Table 5.1: Classification Report

5.3 Performance of Only AffectNet v2 Dataset:

Here, we only used the AffectNetv2 dataset, which has 26,171 image files. All the image sizes were 96*96.The original size of the image was 96*96.The output layer of the EfficientB0 Model was modified to accommodate seven classes, corresponding to the severity levels of emotion. The model was trained for only 10 epochs because after that the validation accuracy was dropping, and it shows the training accuracy of 98.7% and validation accuracy of 70.54%. The resulting training progress is visualized in the figure below.

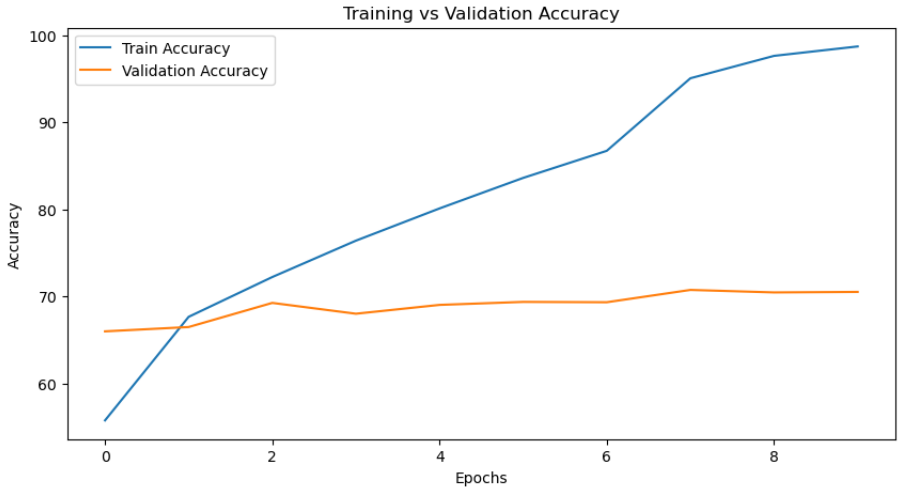


Figure 5.5: Training vs Validation Accuracy

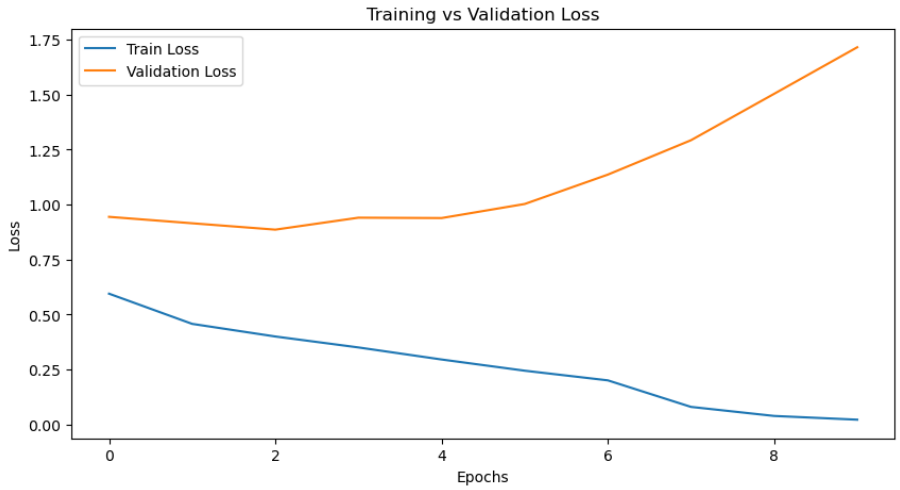


Figure 5.6: Training vs Validation Loss

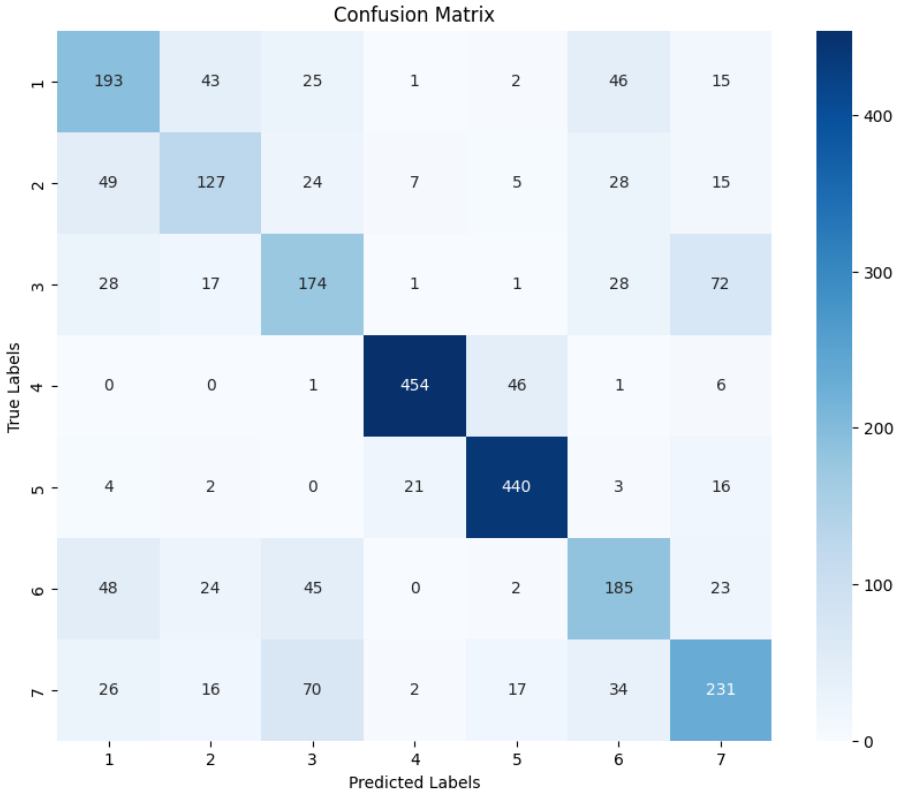


Figure 5.7: Confusion Matrix

Class	Precision	Recall	F1-score	Support
Anger	0.55	0.59	0.57	325
Disgust	0.55	0.50	0.52	255
Fear	0.51	0.54	0.53	321
Happy	0.93	0.89	0.91	508
Neutral	0.86	0.91	0.88	486
Sad	0.57	0.57	0.57	327
Surprise	0.61	0.58	0.60	396
Accuracy	0.69			2618
Macro avg	0.66	0.65	0.65	2618
Weighted avg	0.69	0.69	0.69	2618

Table 5.2: Classification Report

5.4 Performance of Our Model:

We combine the both FER 2013 and Affectnet v2 dataset.After implementing our dataset in our Best_FER Efficient B0 model, it shows the training accuracy of 85.92% and validation accuracy of 63.69%.Our model performs better if we compare the pretrain model ResNet-50 using the combine dataset. The ResNet-50 model achieves a validation accuracy of 43.36%, but our model attains a validation accuracy of 63.69%, much surpassing ResNet-50. Our model effectively captures hierarchical

spatial patterns such as edges, textures, and complex features through its convolutional layers. In order to rectify the imbalance, we determined the class weights by analysing the frequency of each class in the dataset. Class weighting is designed to elevate the significance of minority classes, such as "disgust," to ensure that the model places a greater emphasis on them during the training process. Additionally, the use of dilation rates in the convolutional layers enables the model to capture a larger receptive field without increasing computational cost, facilitating the learning of broader patterns efficiently.Also, with adequate dataset preprocessing, we were able to get this outcome. As we continue to examine the shortcomings, we hope to increase the accuracy of this dataset on the Best_FER model. As we can see in the figure, both the training and validation accuracy increase over time.

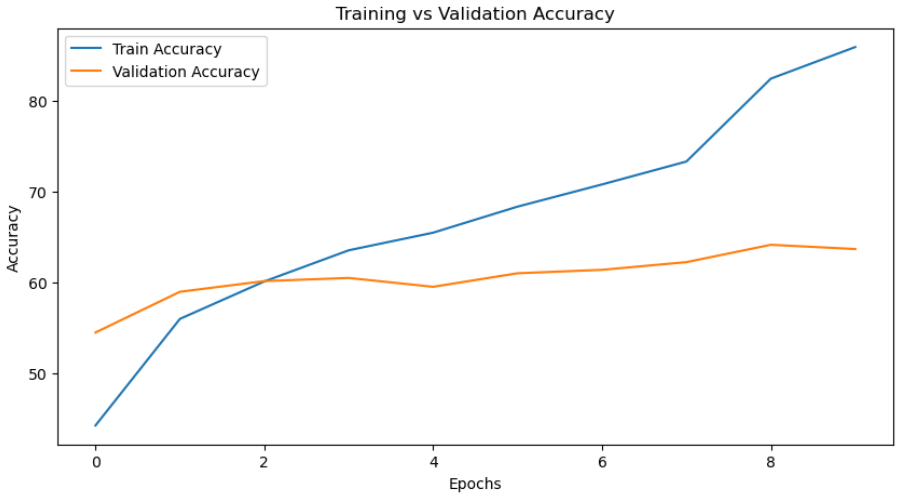


Figure 5.8: Training vs Validation Accuracy

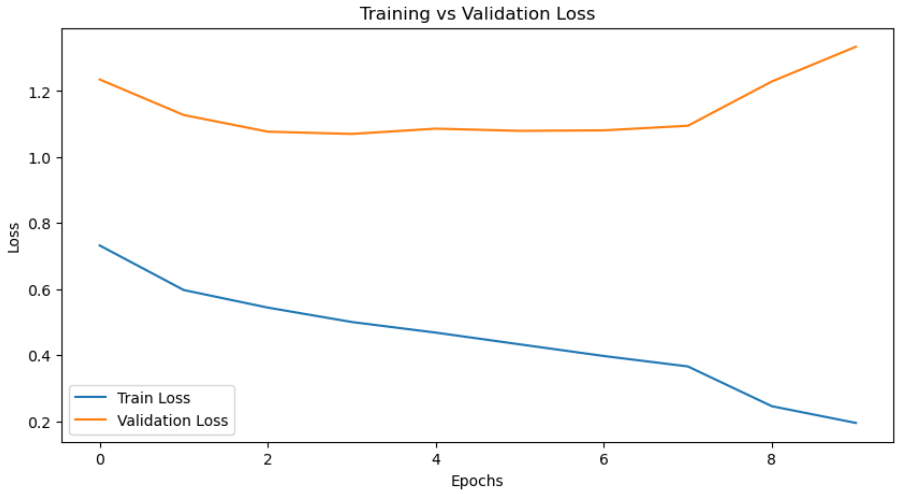


Figure 5.9: Training vs Validation Loss

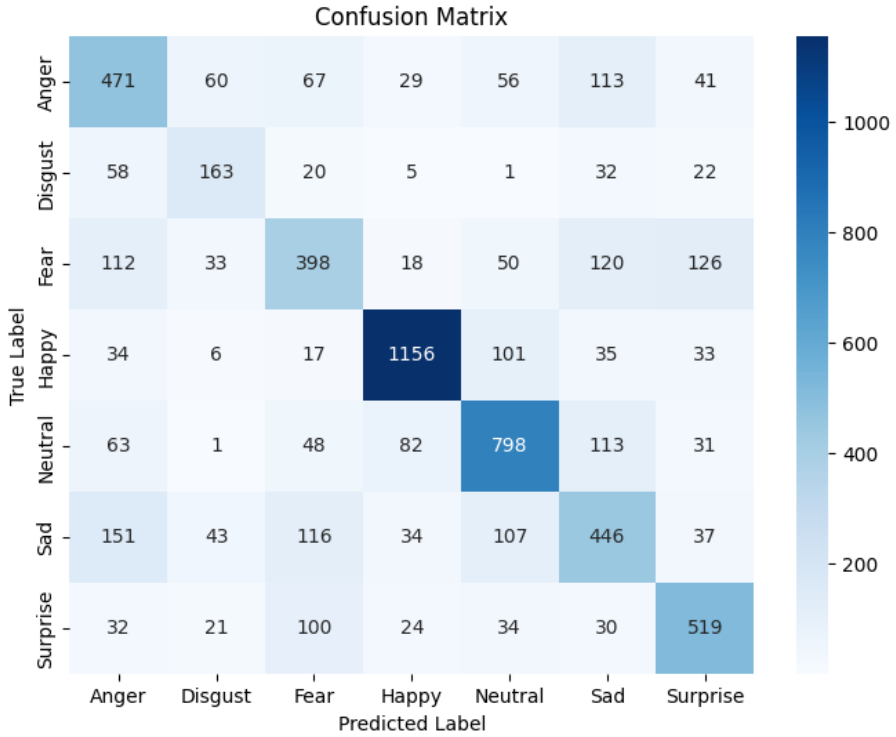


Figure 5.10: Confusion Matrix

Class	**Precision**	**Recall**	**F1-score**	**Support**
Anger	0.51	0.56	0.54	837
Disgust	0.50	0.54	0.52	301
Fear	0.52	0.46	0.49	857
Happy	0.86	0.84	0.85	1382
Neutral	0.70	0.70	0.70	1136
Sad	0.50	0.48	0.49	934
Surprise	0.64	0.68	0.66	760
Accuracy	0.64			6207
Macro avg	0.60	0.61	0.61	6207
Weighted avg	0.64	0.64	0.64	6207

Table 5.3: Classification Report

The classification performance metrics for the model, including precision, recall, F1-score, and support, for each emotion class are presented below. The model achieves an overall accuracy of 64%, indicating moderate performance across all classes. Precision values range from 50% to 86%, reflecting the model's ability to correctly identify positive instances for each class while minimizing false positives. The highest precision is observed for the happy class (86%), while other classes, such as disgust and sad, have lower precision values around 50%. Recall values range from 46% to 84%, demonstrating the model's ability to correctly capture most positive instances without missing true positives. The happy class has the highest recall (84%), while the fear class shows a lower recall of 46%, suggesting some misclassifications. F1-score values, which represent the harmonic mean of precision and recall, range from 49% to 85%, indicating a balance between the two

metrics. The happy class maintains the highest F1-score (85%), while the fear and sad classes have lower F1-scores around 49%. Support represents the number of actual instances for each class, showing a relatively balanced dataset, with the happy and neutral classes having higher representation, while the disgust class has the lowest count. Overall, the macro and weighted averages confirm consistent but moderate performance across all classes, highlighting areas for improvement, particularly in underrepresented emotions such as disgust and fear.

5.4.1 Class wise performance :

- **Class 0 (Anger):** The model achieved a precision of 51%, recall of 56%, and F1-score of 54%, indicating moderate ability to correctly identify anger while still misclassifying some instances.
- **Class 1 (Disgust):** The model achieved a precision of 50%, recall of 54%, and F1-score of 52%, showing that disgust is one of the more challenging emotions to classify accurately.
- **Class 2 (Fear):** With a precision of 52%, recall of 46%, and F1-score of 49%, the model struggles slightly with fear, missing a significant number of true instances.
- **Class 3 (Happy):** The model performs best in this category, achieving precision of 86%, recall of 84%, and F1-score of 85%, demonstrating strong performance in recognizing happiness.
- **Class 4 (Neutral):** The model obtained a precision of 70%, recall of 70%, and F1-score of 70%, indicating balanced performance in detecting neutral expressions.
- **Class 5 (Sad):** The model achieved a precision of 50%, recall of 48%, and F1-score of 49%, suggesting some difficulty in distinguishing sadness from other emotions.
- **Class 6 (Surprise):** With a precision of 64%, recall of 68%, and F1-score of 66%, the model performs relatively well in recognizing surprise compared to other lower-performing categories.

Overall, the model demonstrates strong performance for the happy and neutral classes, while disgust, fear, and sad classes show room for improvement. The macro and weighted averages suggest consistent but moderate classification accuracy across all emotion categories.

5.4.2 Comparison Between Our Model and Pre-trained (ResNet-50) Model:

Model	Training Accuracy (%)	Validation Accuracy (%)
Our Model	85.92	63.69
ResNet-50 (Pre-trained)	68.93	43.36

Table 5.4: Comparison Between Our Model and ResNet-50 Model

5.4.3 Feedback Generation:

After completing 50 epochs for sequence generation using LSTM with hyperparameter optimization, the model achieved accuracy of 83.18% and a loss value of 0.6493. The Train accuracy graph indicates the performance of the model

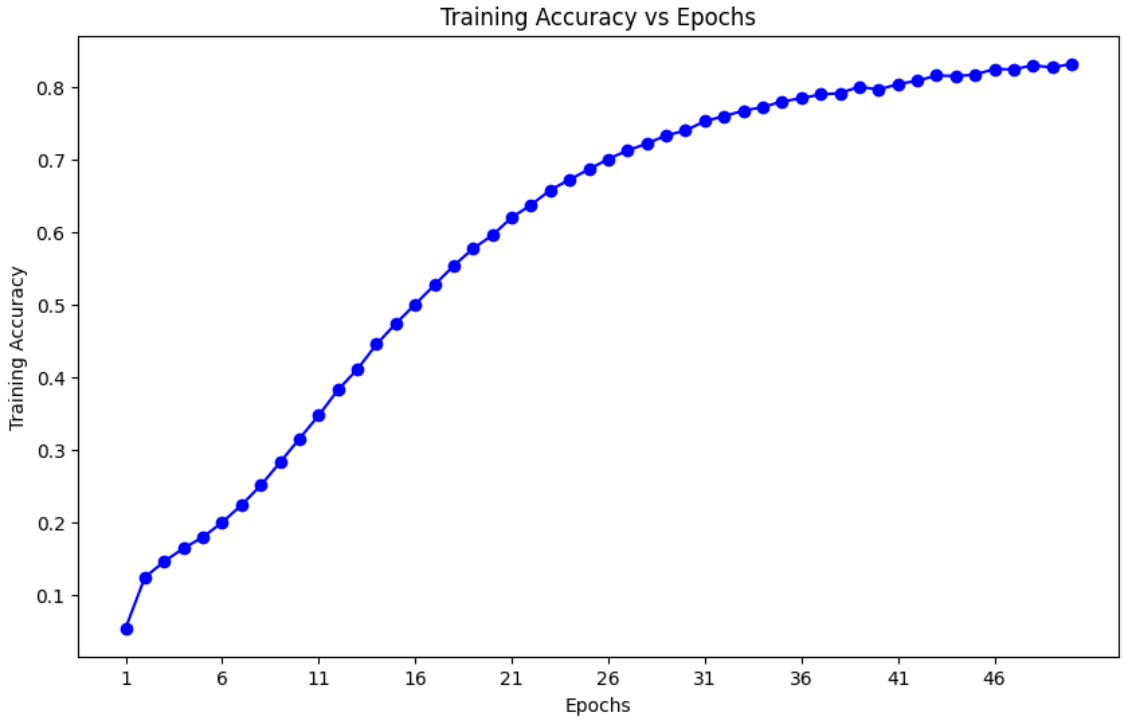


Figure 5.11: Training Accuracy vs Epochs.

Chapter 6

Performance Analysis

The performance of our real-time facial expression recognition (FER) system, integrated with Bengali audio feedback, was evaluated using key classification metrics such as accuracy, precision, recall, and F1-score. Our model was trained on a combined dataset of FER 2013 and AffectNet V2, and its performance was assessed using the Best_FER Efficient B0 model.Our model performs better than pre train model like ResNet-50. The ResNet-50 model has a validation accuracy of 43.36%, while our model has a 63.69% validation accuracy, which is much higher than ResNet-50. After training, our model achieved a training accuracy of 85.92% and a validation accuracy of 63.69%, demonstrating its ability to capture hierarchical spatial features such as edges, textures, and complex patterns. In order to rectify the imbalance, we determined the class weights by analysing the frequency of each class in the dataset. Class weighting is designed to elevate the significance of minority classes, such as "disgust," to ensure that the model places a greater emphasis on them during the training process.Balancing the dataset, together with estimated class weights, helps to reduce overfitting and underfitting. Furthermore, dilated convolutions were employed to expand the receptive field without increasing computational cost, facilitating better feature extraction across varying facial expressions. A detailed classification performance analysis shows an overall accuracy of 64%, with precision, recall, and F1-score values varying across different emotion classes. The happy emotion achieved the best classification performance with 86% precision, 84% recall, and an F1-score of 85%, whereas the disgust and fear categories had lower scores, indicating challenges in distinguishing these emotions.

Class	**Precision**	**Recall**	**F1-score**
Anger	51%	56%	54%
Disgust	50%	54%	52%
Fear	52%	46%	49%
Happy	86%	84%	85%
Neutral	70%	70%	70%
Sad	50%	48%	49%
Surprise	64%	68%	66%

Table 6.1: Class-wise Performance Summary

The macro and weighted averages indicate a moderate yet consistent performance across all classes. While the model performs well in recognizing happy, neutral, and

surprise expressions, there is room for improvement in detecting emotions like disgust, fear, and sadness. Future enhancements, such as advanced augmentation techniques, fine-tuning hyperparameters, and leveraging attention mechanisms, could further improve the classification performance, particularly for underrepresented classes.

Feedback model, developed with LSTM, achieved an accuracy of 83.18% in 50 epochs, and post-processing with a speech-to-text system in Bengali increased audio output naturality and fluency. Optimized for real-time, the custom model achieved lesser inference time over deep pre-trained networks, reduced model footprint through model pruning and quantization, and optimized training approaches for generalizability over a range of face datasets. The proposed model for FER exhibited high accuracy and strong performance over pre-trained networks and can be utilized for real-time use. The audio feedback system in Bengali included an additional level of access, with effective emotion feedback and feedback generation. In the future, work will extend towards improving model adaptability over changing lights, occlusions, and cultural diversity in face expression.

Chapter 7

Conclusion

This research proposes a real-time face expression recognition (FER) system with feedback in Bengali, offering a powerful emotion detection and feedback improvement mechanism. With deep learning approaches, including hybrid CNN-RNN architectures and feedback with LSTM, emotion is detected accurately and feedback generated in audio form in Bengali language.

Comparisons between the in-house and pre-trained models both attested to the Best_FER model's supremacy with high accuracy and generalizability over a range of datasets. The experiments reiterate optimized neural architectures, regularization, and flexible learning techniques in delivering high-performance emotion perception.Integration with audio feedback in Bengali language extends the usability of systems, particularly for native speakers, and enables rich interfaces between humans and machines.The real-time feature enables the system to utilize its capabilities in a variety of applications, including assistive technology, educational software, and interfaces for customer service.Although work holds a lot of promise, adaptations in terms of changing lights, occlusions, and variation in face expression in terms of cultures present will have improvements in future work.Robust model, incorporation of larger and larger datasets, and real-time processing capabilities will have improvements in future work.

The use of multimodal techniques, including speech and bio-data, could enable even enriched emotion perception.The work concludes with a contribution towards enriching the field of affective computation through an efficient and rich face expression recognition system with feedback in Bengali and sets a basis for future work in multimodal deep learning, paving a path towards even more naturally and emotionally rich interfaces between humans and computers.

Bibliography

[1] D. Jurafsky, *Speech and language processing*, 2000.

[2] S. Baron-Cohen, S. Wheelwright, J. Hill, Y. Raste, and I. Plumb, "The "reading the mind in the eyes" test revised version: A study with normal adults, and adults with asperger syndrome or high-functioning autism," *Journal of child psychology and psychiatry*, vol. 42, no. 2, pp. 241–251, 2001.

[3] M. J. Scherer, *Assistive technology: Matching device and consumer for successful rehabilitation.* American Psychological Association, 2002.

[4] D. Goleman, *Emotional intelligence: Why it can matter more than IQ*. Bantam, 2005.

[5] S. A. Hossain, M. L. Rahman, and F. Ahmed, "A review on bangla phoneme production and perception for computational approaches," in *Proceedings of the 7th WSEAS International Conference on Mathematical Methods and Computational Techniques In Electrical Engineering*, 2005, pp. 346–354.

[6] C. D. Manning, P. Raghavan, and H. Schütze, "Boolean retrieval," *Introduction to information retrieval*, pp. 1–18, 2008.

[7] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[8] G. Hofstede, "Dimensionalizing cultures: The hofstede model in context," *Online readings in psychology and culture*, vol. 2, no. 1, p. 8, 2011.

[9] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 689–696.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[11] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[12] A. Karpathy and F. Li, "Deep visual-semantic alignments for generating image descriptions," Jun. 2015, pp. 3128–3137. DOI: 10.1109/CVPR.2015.7298932.

[13] F. Chollet, "The limitations of deep learning," *Deep learning with Python*, 2017.

[14] L. Perez, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*, 2017.

[15] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[16] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural networks*, vol. 106, pp. 249–259, 2018.

[17] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[18] L. N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay," *arXiv preprint arXiv:1803.09820*, 2018.

[19] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019. DOI: 10.1186/s40537-019-0197-0.

[20] K. N and A. Patil, "Multimodal emotion recognition using cross-modal attention and 1d convolutional neural networks," in *Interspeech*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:226202203.

[21] M. Rahman, M. S. Hossain, A. M. Sadek, and G. Muhammad, "Multimodal deep learning approach for emotion recognition," *IEEE Access*, vol. 8, pp. 38 984–38 994, 2020. DOI: 10.1109/ACCESS.2020.2976114.

[22] S. Nadimuthu, K. R. Dhanaraj, and J. kanthan, "Brain tumor classification using convolution neural network," *Journal of Physics: Conference Series*, vol. 1916, p. 012 206, May 2021. DOI: 10.1088/1742-6596/1916/1/012206.

[23] M. Maithri, U. Raghavendra, A. Gudigar, *et al.*, "Automated emotion recognition: Current trends and future perspectives," *Computer Methods and Programs in Biomedicine*, vol. 215, p. 106 646, 2022, ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2022.106646. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260722000311.

[24] S. Monica and R. Roseline Mary, "Face and emotion recognition from real-time facial expressions using deep learning algorithms," in *Congress on Intelligent Systems*, M. Saraswat, H. Sharma, K. Balachandran, J. H. Kim, and J. C. Bansal, Eds., Singapore: Springer Nature Singapore, 2022, pp. 451–460, ISBN: 978-981-16-9416-5.

[25] Y. Wang, W. Song, W. Tao, *et al.*, *A systematic review on affective computing: Emotion models, databases, and recent advances*, 2022. arXiv: 2203.06935 [cs.MM].

[26] R. Venkatakrishnan, M. Goodarzi, and M. A. Canbaz, "Exploring large language models' emotion detection abilities: Use cases from the middle east," in *2023 IEEE Conference on Artificial Intelligence (CAI)*, 2023, pp. 241–244. DOI: 10.1109/CAI54212.2023.00110.

[27] Y. Bian, D. Küster, H. Liu, and E. G. Krumhuber, "Understanding naturalistic facial expressions with deep learning and multimodal large language models," *Sensors*, vol. 24, no. 1, 2024, ISSN: 1424-8220. DOI: 10.3390/s24010126. [Online]. Available: https://www.mdpi.com/1424-8220/24/1/126.

[28] D. Li, X. Liu, B. Xing, *et al.*, *Eald-mllm: Emotion analysis in long-sequential and de-identity videos with multi-modal large language model*, 2024. arXiv: 2405.00574 [cs.CV].

[29] S. Campeau, W. Falls, and W. Cullinan, *Picard, rw (1997), affective computing. cambridge.*