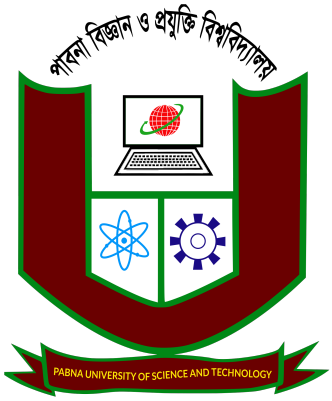
Advancing Bangla Speech Emotion Recognition with Emoformer and Self-Attention Mechanisms



Thesis

Course Code: ICE - 6000

M.Sc. (Engineering) Examination - 2021

A thesis paper submitted to the Department of Information and Communication Engineering, Pabna University of Science and Technology in partial fulfillment of the requirements for the degree of Master of Science in Engineering in Information and Communication Engineering

Prepared by:

MST. ASMA KHATUN

Roll No: 21040624

Reg. No: 1065220

Session: 2020-2021

Department of Information and Communication Engineering

Pabna University of Science and Technology

Pabna-6600, Bangladesh

January 2026

Certificate of the Supervisor

It is certified that the research work incorporated in this thesis entitled **"Advancing Bangla Speech Emotion Recogni- tion with Emoformer and Self-Attention Mechanisms"** is the original work carried out by **MST. ASMA KHATUN** under my supervision, and it fulfills the conditions laid out in the Pabna University of Science and Technology Ordinances. The research work included in this thesis forms a distinct contribution to knowledge. The thesis contains work worthy of consideration for the award of the degree of Master of Science in Engineering in Information and Communication Engineering (ICE).

.....

Dr. Md. Sarwar Hosain
Professor
Department of Information and Communication Engineering
Pabna University of Science and Technology, Pabna-6600, Bangladesh.

Declaration

This is to certify that the thesis entitled **"Advancing Bangla Speech Emotion Recognition with Emoformer and Self-Attention Mechanisms"**, has been carried out by me under the course entitled **"Thesis/Project (ICE-6000)"**. I also declare that this work has not been submitted elsewhere in part or full for the requirement of any degree or for any other purpose, except for academic publication.

.....

MST. ASMA KHATUN
Roll No: 21040624
Reg. No: 1065220
Department of Information and Communication Engineering
Pabna University of Science and Technology, Pabna-6600, Bangladesh.

Certificate of the Supervisor

It is certified that the research work incorporated in this thesis entitled **"Advancing Bangla Speech Emotion Recogni- tion with Emoformer and Self-Attention Mechanisms"** is the original work carried out by **MST. ASMA KHATUN** under my supervision, and it fulfills the conditions laid out in the Pabna University of Science and Technology Ordinances. The research work included in this thesis forms a distinct contribution to knowledge. The thesis contains work worthy of consideration for the award of the degree of Master of Science in Engineering in Information and Communication Engineering (ICE).

.....

Dr. Md. Sarwar Hosain
Professor
Department of Information and Communication Engineering
Pabna University of Science and Technology, Pabna-6600, Bangladesh.

Certificate of the Co-supervisor

It is certified that the research work incorporated in this thesis entitled **"Advancing Bangla Speech Emotion Recogni- tion with Emoformer and Self-Attention Mechanisms"** is the original work carried out by **MST. ASMA KHATUN** under my supervision, and it fulfills the conditions laid out in the Pabna University of Science and Technology Ordinances. The research work included in this thesis forms a distinct contribution to knowledge. The thesis contains work worthy of consideration for the award of the degree of Master of Science in Engineering in Information and Communication Engineering (ICE).

.....

Dr. Md. Omar Faruk
Professor
Department of Information and Communication Engineering
Pabna University of Science and Technology, Pabna-6600, Bangladesh.

Dedicated to...

My parents

and

My Teachers, whose guidance and support shaped my journey.

Acknowledgements

Firstly, I am sincerely thankful to **Pabna University of Science and Technology** for providing a supportive and intellectually stimulating environment that fostered my academic and personal growth throughout the Master's program.

It is a great honor and privilege to have conducted this project as a part of the **Master of Science in Information and Communication Engineering**, under the **Faculty of Engineering and Technology**, at **Pabna University of Science and Technology**.

I want to extend my deepest gratitude to my respected supervisor for his constant guidance, encouragement, and insightful advice throughout the course of this work. His unwavering support and belief in my potential have been instrumental in completing this research successfully.

Finally, I am genuinely thankful to everyone who supported me throughout this journey: my teachers, friends, and family. Your motivation, assistance, and kind words have played a significant role, and I am truly grateful for all the encouragement I have received.

The Author

Abstract

Speech Emotion Recognition (SER) plays a critical role in enhancing human computer interaction, supporting mental health assessment, and enabling adaptive and personalized intelligent systems. Although Bangla is the seventh most widely spoken language globally, it remains underrepresented in speech emotion research, primarily due to limited annotated datasets and linguistic diversity. Addressing this gap, this study proposes a novel Bangla speech emotion recognition framework based on an attention driven Emoformer architecture, which collaboratively integrates convolutional neural networks with transformer encoders to effectively model both local acoustic patterns and long-range temporal dependencies.

The proposed system utilizes a hybrid feature representation that combines handcrafted acoustic descriptors, such as Mel-Frequency Cepstral Coefficients (MFCCs), with deep speaker-agnostic embeddings, namely X-vectors. This dual-feature strategy enables the model to capture complementary emotional cues related to spectral characteristics, prosody, and speaker variability, thereby improving robustness across diverse speaking styles. Extensive experiments are conducted on the BanglaSER dataset, comprising 1,467 labeled speech utterances collected from 34 speakers expressing five emotional states: angry, happy, sad, surprise, and neutral.

Experimental results demonstrate that the proposed Emoformer model achieves an overall classification accuracy of 86%, outperforming existing state-of-the-art approaches for Bangla SER. The multi-head self-attention mechanism allows the model to selectively emphasize emotionally salient temporal and spectral regions of speech, effectively addressing challenges arising from Bangla's rich phonetic structure and dialectal variations. Notably, the system exhibits exceptional performance in recognizing neutral emotion, achieving a recall of 1.00 and an area under the curve (AUC) of 0.99, while maintaining strong and consistent performance across all remaining emotion categories, with F1-scores exceeding 0.75. These findings establish a new benchmark for Bangla speech emotion recognition and demonstrate the effectiveness of hybrid CNN Transformer architectures in low-resource language settings. The proposed approach not only advances the state of Bangla SER but also offers a scalable and transferable framework for emotion recognition in other underexplored languages.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Speech-based emotion recognition (SER) is used in natural language processing and human–computer interaction (HCI), among other fields [1]. SER can improve HCI systems by supporting more personalized and human-like interactions. It is valuable in various fields such as marketing, education, mental health, speech synthesis, and customer satisfaction [1]. For example, SER can improve the overall user experience by identifying dissatisfied users and providing insights into their preferences and behavior. To assess speech emotions, a range of methods and strategies are employed, such as machine learning algorithms along with statistical and probabilistic models [2]. Deep learning methods have recently become key in this area [3, 4]. The use of deep learning for speech emotion recognition has demonstrated promising results, with approaches such as CNNs [5], DBNs, RNNs, and LSTMs [6]. Limited research has been done on detecting emotions in the Bangla language, highlighting the need for a speech-emotion recognition system for Bangla.

Azmin et al. [7] demonstrated that Multinomial Naïve Bayes, combined with features such as stemming, POS tagging, n-grams, and tf-idf, can accurately categorize Bangla text into three emotion classes: happy, sad, and angry, achieving an accuracy of 78.6%. Their findings suggest that traditional machine learning techniques remain effective even with limited Bangla linguistic resources.

Badhon et al. [8] highlighted the increasing significance of natural language processing in enhancing communication between humans and intelligent systems, especially through spoken interaction. Their research pointed out that although English has benefited from extensive research and commercial progress in speech recognition,

Bangla, despite being the eighth most spoken language worldwide with around 163 million speakers, remains underrepresented. The authors reviewed current efforts in Bangla speech recognition, illustrating various research attempts to tackle speech processing issues using different approaches.

Ryhan et al. [9] created two deep learning models, BiGRU and CNN-BiLSTM, to identify emotions in Bangla text across six categories. Their evaluation using accuracy, precision, recall, and F1-score showed that neural network architectures can effectively understand context and perform well, especially on datasets translated with Google Translator. The results suggest that deep learning methods have strong potential for Bangla emotion detection, surpassing traditional machine learning techniques.

Das et al. [10] introduced encoder-decoder models, such as attention, LSTM, and GRU decoders, to categorize Bengali social media comments into seven hate speech types. The attention-based decoder achieved the highest accuracy of 77%. The research shows that encoder-decoder structures can effectively identify contextual features in Bangla text. However, the study is limited by a small dataset, focusing only on Facebook comments, and moderate accuracy, suggesting the need for larger, more varied datasets and enhanced model performance.

Purba et al. [11] created a new Bangla document dataset labeled with three emotions: Happy, Sad, and Angry, and used feature extraction techniques like Bag of Words and Word Embedding. They tested various classifiers, including Logistic Regression, Multinomial Naïve Bayes, ANN, and CNN. The Multinomial Naïve Bayes classifier achieved the highest accuracy at 68.27%, indicating that traditional machine learning methods can perform reasonably well for Bangla emotion detection. However, the study's limitations include a small dataset, only three emotion categories, and moderate accuracy, emphasizing the need for larger, more diverse datasets and investigation of more advanced models.

Midhra et al. [12] examined the current landscape of Bengali Automatic Speech Recognition (ASR) systems, pointing out various language-dependent and independent hurdles in creating precise models. Their research underscores that ASR architectures must be tailored to fit Bengali's grammatical and phonetic features. Nevertheless, Bengali ASR research remains in its infancy, characterized by scarce resources, tools, and solid implementations, highlighting the need for concentrated efforts in developing language-specific ASR systems.

Das et al. [13] created a Bengali emotion dataset containing 6,243 texts and employed machine learning, deep learning, and transformer-based methods to identify six core emotions: anger, fear, disgust, sadness, joy, and surprise. Their results showed

that transformer models, especially XLM-R, surpassed other techniques by achieving the highest weighted F1-score, highlighting the strength of transformer architectures in Bangla emotion classification. Nevertheless, the study's limitations include a relatively small dataset and a focus on only six emotion categories, which suggests the need for larger, more varied datasets and further testing of transformer models in real-world scenarios.

Ali et al. [14] created BanglaSenti, a lexicon-based dataset with 61,582 Bangla words labeled for positive, negative, and neutral sentiments, designed for sentiment analysis and adaptable for emotion detection. Their research included model simulations to illustrate the dataset's usability, highlighting its potential for BNLP tasks like opinion mining and review analysis. Nonetheless, the study has limitations due to its lexicon-only approach, which may miss important contextual nuances in complex sentences. Additionally, it predominantly addresses polarity rather than detailed multi-class emotion classification, underscoring the need for more comprehensive datasets and models to achieve nuanced emotion detection in Bangla.

Das et al. [15] introduced BEmoD, a Bengali emotion dataset with 5,200 texts categorized into six basic emotions: anger, fear, surprise, sadness, joy, and disgust. The dataset was created through data crawling, pre-processing, labeling, and verification, resulting in a high level of annotation agreement with a Cohen's K score of 0.920. Although BEmoD is a valuable resource for Bengali emotion analysis, its relatively small size and focus on only six basic emotions highlight the need for larger and more diverse datasets to enhance model training and generalization.

Ahmed et al. [16] created a CNN-based system for recognizing Bangla hand signs, designed to help people with speech disabilities by translating hand signs into spoken Bangla. The system reached a 92% accuracy on validation data, showing its effectiveness and potential for real-world communication support. However, the study only focused on hand sign digits and did not include a wider variety of gestures or continuous sign language, highlighting the need for more detailed datasets and broader models for complete Bangla sign language recognition.

Our primary objective is to classify speech emotions using an attention-based Emoformer model. Recognizing emotions in spoken language, especially in Bangla, is challenging due to diverse linguistic usage, social and cultural influences, personal experiences, and limited existing evidence [17]. Again, individual and cultural differences, along with the range of emotional expressions in tones, dialects, and speech rates, pose significant challenges for algorithms attempting to detect emotions in Bangla speech [18]. Recently, many researchers have used traditional handcrafted feature-based

machine learning methods, including MFCCs, Chroma, and Spectral Contrast, to identify emotions in Bangla speech [6].

However, their accuracy performance remains unsatisfactory. Recently, some researchers have been using deep learning methods such as CNN, LSTM, and BiLSTM models to enhance performance accuracy [19]. However, the research still encounters difficulties in attaining high accuracy and robust generalization because of insufficient effective features. To address this, we proposed combining handcrafted features with deep learning features to create a Bangla speech emotion recognition system.

1.2 Application

Speech Emotion Recognition (SER) is an expanding field of research dedicated to automatically detecting human emotions through speech signals. By analyzing emotional cues within voice, SER allows intelligent systems to engage with humans more naturally, adaptively, and effectively.

- **Human–Machine Interaction:** SER greatly improves human-computer interaction by enabling machines to recognise and react to users' emotions. Emotion-aware virtual assistants, chatbots, and smart devices can modify their tone, responses, and behavior depending on whether a user appears happy, frustrated, or stressed. This results in more natural, empathetic, and engaging interactions, enhancing user satisfaction and the usability of intelligent systems.

- **Healthcare and Mental Well-being:** In healthcare, especially mental health care, SER can act as a non-invasive method to track emotional and psychological states. Regular analysis of speech patterns can identify early indicators of stress, depression, anxiety, or emotional instability. These systems assist clinicians by offering objective emotional data, supporting early diagnosis, remote tracking, and prompt intervention, particularly in telemedicine and long-term patient management.

- **Customer Experience Management:** SER is essential in customer service and call centres for detecting emotions like anger, frustration, or dissatisfaction instantly. This enables organizations to prioritize urgent calls, direct customers to capable agents, and modify service approaches as needed. Recognising customer emotions helps companies enhance service quality, decrease customer churn, and strengthen customer relationships.

- **Education and Learning Systems:** In educational environments, SER supports emotion-aware tutoring and e-learning platforms that adjust teaching strategies according to students' emotional responses. By identifying feelings like boredom, confusion, or frustration via speech analysis, these intelligent systems can alter content difficulty, supply extra explanations, or give encouragement. This tailored method boosts student motivation, improves learning effectiveness, and elevates overall academic results.

- **Entertainment and Media:** SER enhances immersive and interactive experiences in entertainment, gaming, and multimedia applications. Games and virtual environments can dynamically tailor their responses to players' emotional states by modifying difficulty, storylines, or background features. Likewise, emotion-aware media systems can suggest or modify content according to users' moods, increasing engagement and enjoyment.

- **Security and Surveillance:** In security contexts, SER helps identify stress, fear, or deception in voice authentication and surveillance. Analysing speech emotions can strengthen traditional biometric systems by providing an extra security layer. This is especially valuable in high-risk settings like border control, emergency responses, and sensitive access points, where emotional signals might reveal threats or unusual activity.

1.3 Motivation of Research

Although Bangla is the seventh most spoken language globally, with over 230 million speakers, research on emotion recognition in this language is limited compared to high-resource languages like English, Mandarin, and German. Bangla's rich acoustic diversity, including unique phonemes, varied intonational patterns, and significant dialectal differences, offers both challenges and opportunities for emotion recognition. The cultural and linguistic nuances of Bangla emotional expression differ greatly from Western languages, requiring dedicated research rather than relying solely on cross-lingual transfer learning. Additionally, as voice-based technologies become more common in Bangladesh and Bengali-speaking regions, there is an urgent need for emotionally intelligent systems that understand and respond in native language. Current speech emotion recognition (SER) systems, mainly trained on English or other resource-rich languages, do not capture the specific acoustic-prosodic features of Bangla emotional speech, leading to poor performance in real-world applications. This research aims to address this

gap and ensure Bengali speakers benefit equally from advances in affective computing and emotionally aware AI systems.

1.4 Research Gap

Current studies on Bangla speech emotion recognition highlight several important gaps in the research.

- **Limited Availability and Quality of Datasets:** A major challenge in Bangla SER research is the shortage of large, high-quality emotional speech datasets. Current Bangla corpora tend to be small, imbalanced in emotion classes, and recorded in controlled or artificial settings. They often lack diversity in speaker age, gender, dialect, and speaking style, as well as real-world acoustic variability. This limits effective model training, increases overfitting risks, and makes it difficult to compare results across studies, thereby slowing progress toward developing generalized and dependable SER systems.

- **Over-Reliance on Handcrafted Acoustic Features:** Most previous research on Bangla SER primarily relies on traditional handcrafted features such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch, energy, and formants. These features are often used with standard machine learning classifiers like Support Vector Machines (SVM) and Gaussian Mixture Models (GMM). Although these methods establish baseline performance, they struggle to effectively capture complex, non-linear emotional patterns in speech, leading to limited accuracy and poor generalization across different speakers and recording environments.

- **Limited Effectiveness of Traditional Machine Learning Models:** Traditional classifiers employed in Bangla SER studies often face challenges with high-dimensional emotional features and variability between speakers. These models demand significant feature engineering and manual adjustments but typically fall short in capturing the temporal and contextual aspects of emotional speech. Consequently, their performance lags behind that of advanced deep learning methods used in high-resource languages.

- **Lack of Attention Mechanism Utilization:** Current Bangla SER models seldom include attention mechanisms that can dynamically emphasize emotionally important parts of speech while reducing focus on irrelevant or neutral segments. Attention-based models have shown significant performance gains in high-resource

language SER by enabling networks to concentrate on key emotional cues. The scarce investigation of these mechanisms in Bangla SER highlights a notable research gap and a valuable opportunity for improving performance.

- **Low Recognition Accuracy and Emotion Confusion:** Most Bangla SER systems currently report accuracy below 85%, significantly lower than the state-of-the-art results in resource-rich languages. Moreover, models often find it difficult to differentiate between acoustically similar emotions like anger and excitement or sadness and boredom. This ongoing confusion underscores the limitations of current methods and suggests that Bangla SER has not yet achieved mature or competitive performance levels.

1.5 Problem Statement

This research tackles the core issue of insufficient performance and limited accessibility of speech emotion recognition (ser) systems for the Bangla language, even though it ranks among the most widely spoken languages globally. Current SER methods for Bangla encounter three related challenges:

- **Feature representation limitations -** traditional handcrafted features alone are insufficient to capture the complex acoustic-prosodic patterns that define Bangla emotional speech across various speakers and dialects.
- **Modeling inadequacies -** traditional architectures, such as standalone CNNs and RNNs, struggle to effectively capture both local spectral-temporal features and broader contextual dependencies required for precise emotion classification; and
- **Low-resource constraints -** the limited availability of large-scale, diverse, and balanced Bangla emotional speech datasets hinders the training of deep learning models that can effectively generalize to unseen speakers and real-world scenarios.

These challenges lead to less-than-ideal recognition accuracy, usually below 85%, especially when differentiating acoustically similar emotions like angry versus happy or sad versus neutral. Therefore, there's a pressing need for a sophisticated SER architecture that can perform well in low-resource environments and still match the performance of cutting-edge systems designed for high-resource languages.

1.6 Research Objectives

This research mainly aims to achieve the following objectives:

- To create a new hybrid architecture that merges Convolutional Neural Networks with Transformer encoders enhanced by multi-head self-attention, specifically aimed at capturing both local spectral-temporal features and global contextual dependencies in Bangla emotional speech.
- To develop a dual feature representation strategy that combines Mel-Frequency Cepstral Coefficients (MFCCs) for spectral-temporal features and X-vectors for speaker-independent prosodic and emotional cues, aiming to maximize discriminative information extraction from limited training data.
- To utilize multi-head self-attention mechanisms with multiple attention heads, the model can concurrently focus on various emotionally relevant aspects of speech. This enables dynamic weighting of key segments and features, leading to better emotion discrimination.
- To reach state-of-the-art results on the BanglaSER dataset, outperforming current methods by thoroughly evaluating multiple metrics such as accuracy, precision, recall, F1-score, and AUC across all emotion categories (angry, happy, sad, surprise, neutral).
- To develop a solid framework for speech emotion recognition in low-resource language contexts, showcasing how advanced attention-based architectures can effectively generalize from limited training data and offering a replicable approach for other under-resourced languages.
- To perform a thorough performance analysis to identify the model's strengths and weaknesses across various emotion categories, offering insights into specific recognition challenges and potential areas for future research.

1.7 Contribution

This study offers several important advancements in speech emotion recognition, especially for languages with limited resources.

- **Novel Emoformer Architecture:** We present an advanced hybrid deep learning model that integrates CNNs with Transformer encoders and multi-head self-attention, tailored for Bangla speech emotion recognition. This design overcomes previous limitations by effectively capturing local spectral features along with broader temporal information.

- **Dual Feature Integration Strategy:** We suggest a thorough feature engineering approach that integrates handcrafted MFCCs, which capture spectral-temporal features, with deep learning-based X-vectors that encode speaker-independent prosodic patterns. This combination creates a rich, complementary feature set that improves the ability to distinguish emotions.

- **Multi-Head Attention for Emotion Focus:** We employ an eight-head self-attention mechanism that allows the model to focus on multiple emotionally relevant speech features at once, dynamically adjusting the importance of key temporal and spectral aspects. This approach is novel in Bangla SER research.

- **omplete Implementation Pipeline:** We provide a thorough, reproducible implementation framework that includes data preprocessing steps, two feature extraction methods, detailed architectural details, and training settings. This framework can be easily adapted for similar SER tasks or extended to other low-resource languages, contributing to the advancement of affective computing in underserved linguistic communities.

The remainder of this paper is structured as follows: Section 2 covers related work on Bangla SER and transformer architectures; Section 3 analyses attention mechanisms, ablation studies, and dataset preprocessing; Section 4 discusses experimental setup and results; and Section 5 concludes with potential future directions.

Chapter 2

Literature Review

2.1 Methodology Review

Speech Emotion Recognition (SER) has attracted significant interest recently, especially for Bangla-language applications, due to its crucial role in human–computer interaction, mental health monitoring, and voice-driven systems. Current studies show notable progress in deep learning techniques, dataset development, feature extraction methods, and cross-lingual assessments. Early works mainly relied on traditional machine learning techniques using acoustic features such as MFCC, achieving moderate performance but struggling with complex emotional variations [20]. With advances in deep learning, hybrid architectures such as DCNN-BLSTM have been introduced, significantly improving performance by capturing both spatial and temporal features of speech signals. The development of datasets such as BanglaSER has further facilitated research by providing structured, balanced emotional speech data [21].

Recent studies highlight the effectiveness of deep learning models, especially CNNs and hybrid approaches, which achieve high accuracy by leveraging features such as log-Mel spectrograms and Zero Crossing Rate [22]. Additionally, combining machine learning and deep learning methods has demonstrated superior performance over traditional approaches [23]. Optimisation techniques such as CNN-LSTM and boosting classifiers have been proposed to improve robustness and address issues like noise sensitivity and class imbalance [24]. Despite these advancements, Bangla SER still faces challenges due to limited data resources and variability in real-world speech [25]. Cross-lingual and language-independent approaches have also been explored, indicating that some prosodic features can generalize across languages, although limitations remain for certain emotions [26].

A prominent method in Bangla SER is the cascaded deep learning model introduced by Billah et al. [27], which utilizes a two-stage classification process for identifying emotions and their intensities. This approach constructs 3D speech signal representations from various 2D transformations and combines CNN, LSTM, and BiLSTM architectures, resulting in excellent results on both the RAVDESS and KBES datasets.

The same authors also emphasize the design and significance of the KBES dataset [28], which includes intensity levels and real-world dialogue-based emotional data, helping to overcome the lack of Bangla emotional corpora. Collectively, these studies highlight the importance of intensity-aware SER modeling and diverse datasets in Bangla.

In addition, other researchers have explored machine learning methods. Sultana et al. [29] show that traditional classifiers like SVM and Random Forest remain competitive when combined with optimized features such as MFCC and statistically derived metrics. Their results highlight that feature selection techniques like correlation analysis and recursive elimination are crucial for enhancing recognition accuracy, especially for neutral and angry emotions. Nevertheless, the reduced performance on sad expressions reveals some limitations of classical machine learning techniques in capturing intricate emotional signals.

Saad et al. [30] investigate the cross-lingual generalization of SER by analyzing emotion recognition in Bangla and English. Their SVM-based assessment shows some level of language independence but highlights cultural factors impacting the recognition of emotions like disgust and fear. Results suggest that native speakers tend to be more accurate in expressing emotions in their own language, which raises concerns about potential dataset bias and speaker familiarity. Dataset development is a crucial aspect of SER research.

Hussain et al. [31] introduce the BanSpEmo corpus, created with non-actors to ensure natural expressions. Their validation showed over 76% correct recognition, indicating reliable human perception. This dataset plays a vital role in Bangla SER by filling the speaker diversity gap. Likewise, Talukder et al. [32] suggest a hybrid CNN-BiLSTM model trained on the SUBESCO dataset, emphasizing the balance between performance and efficiency for IoT device deployment. Although it achieves high accuracy, the absence of real-time testing and reliance on a single dataset raise concerns about potential overfitting. Recent work by Momshad et al. [33] investigates newer architectures such as Wav2Vec2 and ExHuBERT, indicating a trend toward self-supervised and transformer-based feature extraction. Their application of mel-spectrogram images and automatic label-correction tools highlights methodological advancements and results in significant

accuracy gains through label refinement. However, the scalability of transformer models still requires high computational resources.

Deep learning progress is highlighted by Hosain et al. [34], who attain very high accuracy with DNNs utilizing SUBESCO and other datasets. They incorporate data augmentation to mitigate dataset limitations, though the small dataset size restricts how broadly the results can be applied. Despite this, their cross-dataset evaluation shows their model as one of the leading baseline frameworks for Bangla SER.

From a wider regional perspective, Monisha et al. [35] offer an in-depth review of SER research across Indo-Aryan and Dravidian languages, tracing the development from traditional statistical models like KNN and HMM to current neural methods. This situates the Bangla SER field within broader linguistic trends and emphasizes persistent challenges such as limited corpora and cross-cultural differences in emotional expression cues.

Kibria et al. [36] make a valuable contribution to Bangla speech technology research by tackling a key challenge in developing Large Vocabulary Continuous Speech Recognition (LVCSR) systems—specifically, the scarcity of a large, balanced speech corpus for Bangladeshi Bangla. Their work introduces the SUBAK.KO corpus, a new language resource designed to enhance Automatic Speech Recognition (ASR) accuracy for Bangla speakers, considering regional differences. As they point out, the success of modern ASR systems heavily depends on the quality, size, and variety of speech corpora, and historically, Bangla has faced a shortage of such resources.

Hossain et al. [37] present a DCNN-BLSTM architecture and examine its cross-lingual performance with the SUBESCO and RAVDESS datasets. Their findings demonstrate the potential of deep models in emotion recognition across different languages but also expose limitations related to broader cultural and linguistic contexts. Although the model attains promising weighted accuracies of 86.9% and 82.7%, the absence of testing with additional languages and larger datasets restricts its broader applicability. Moreover, challenges related to low-resource languages persist, highlighting the need for scalable SER frameworks capable of adapting across various domains [38].

Traditional machine learning remains pertinent, as demonstrated by Ayon et al. [39], who employ ensemble methods such as RF, DT, KNN, and MLP. Their results show a significant gap between training and testing accuracy, 99% versus 78%, highlighting potential overfitting and sensitivity to the dataset. While ensemble voting enhances robustness, the study emphasizes that dataset size and feature engineering are vital to performance. Nonetheless, the lack of multi-lingual validation and dependence on mixed acted-spontaneous data raises questions about real-world applicability.

Deep learning studies, like Hassan et al. [40], explore CNN-LSTM architectures that achieve strong results on RAVDESS and SUBESCO datasets. Techniques such as AWGN and pitch modification improve the generalization of these models. However, their primary focus is on Bangla and English, and the near-perfect accuracy raises questions about dataset bias and whether the models perform well in varied acoustic settings. Additionally, the lack of comparisons with other models makes it harder to interpret their performance claims.

Biswas et al. [41] highlight the use of deep learning with a CNN-based BSER system trained on an expanded SUBESCO dataset. While achieving a high accuracy of 97.66%, the model may face scalability challenges, especially under real-world noisy conditions. The authors point to future research in hybrid architectures and multimodal fusion, indicating that single-modality systems might not be enough for capturing complex emotional states.

Namey et al. [42] examine hybrid modeling techniques by combining CNN and GRU modules with cochleagram and spectrogram features. Their dual-branch design successfully captures spatial and temporal emotion cues, reaching 92.04% accuracy on BanglaSER. Nonetheless, the study highlights the limitations of using CNNs or LSTMs alone, emphasizing the value of hybrid models. However, it offers limited details on augmentation methods and continues to face challenges due to a lack of diverse datasets.

Begum et al. [43] enhance deep learning methods by incorporating traditional models like SUBESCO and BanglaSER. Their research demonstrates high performance with MFCC-based features and ensemble classifiers, but dependence on acted speech questions real-world applicability. Additionally, the computational demands and use of single-language datasets limit scalability to multilingual and resource-scarce environments.

Saad et al. [17] investigate cross-lingual emotion dynamics and suggest that SER might be partly language-independent. They utilize prosodic features and SVM classifiers to identify similarities and differences between Bangla and English emotions. However, the roles of cultural and environmental factors are still not well-understood, particularly for emotions like disgust and fear. Additionally, more research is needed into how native and non-native speakers express emotions to enhance the effectiveness of practical SER systems.

Sultana et al. [44] examine corpus-driven evaluation by performing human ratings and reliability tests across seven emotions. Their methodology, which uses ICC, ANOVA, and Kappa metrics, confirms the quality of the corpus. However, detecting complex emotions like Disgust continues to be difficult, highlighting the difficulty in recognizing

subtle emotional cues. Additionally, their focus solely on audio limits the potential for broader multimodal emotion research.

Similarly, Roni et al. [45] utilize CNN models to classify seven emotions and their intensity levels, combining MFCC and STFT features with data augmentation. Their findings highlight ongoing misclassification issues among acoustically similar emotions, particularly fear and anger. This points to the importance of developing more discriminative feature extraction methods, possibly through transformers or self-supervised learning.

Finally, Chowdhury et al. [46] highlight the effectiveness of combining CNN and LSTM architectures for feature fusion, attaining high accuracy on SUBESCO and KBES datasets. Their research confirms that hybrid deep learning approaches surpass traditional machine learning, and their comparative evaluations emphasize the significance of dataset choice and preprocessing techniques.

2.2 Accuracy Review

Speech emotion recognition (SER) has made great progress with deep learning techniques [47]. Yet, research on SER in the Bangla language remains limited. A notable early contribution was made by Sadia Sultana, M. Zafar Iqbal, and team in 2021, who developed a deep learning-based SER framework using the Bangla audio-only SUBESCO corpus [48]. Their model integrated deep CNN architectures with bidirectional LSTMs and included a time-distributed flatten (TDF) layer. In cross-lingual and multilingual tests with the SUBESCO and RAVDESS datasets, their TDF-enhanced model outperformed several current CNN-based systems, reaching 86.9% weighted accuracy (WA) on SUBESCO and 82.7% WA on RAVDESS. Earlier, Rahman, Md. Masudur, and colleagues (2018) developed an automatic emotion recognition system for Bengali that relies on speech signals. They utilized MFCCs as static features and MFCC derivatives to capture dynamic aspects. The system employed a support vector machine with an RBF kernel for classification and a modified DTW method for feature matching, attaining an accuracy of 86.08% across 12 speakers [49].

In 2022, Chakraborty et al. proposed a phase-based cepstral feature extraction method utilizing PBCC for speech emotion recognition (SER) [18]. They tested their approach on the SUBESCO and BanglaSER datasets with a gradient boosting machine classifier. The system recorded an average accuracy of about 96% in both speaker-dependent and speaker-independent scenarios, representing a significant advancement over conventional techniques.

In 2020, Dias Issa, M. Fatih Demirci, and collaborators introduced a different approach, developing a 1D CNN-based system that utilizes five key acoustic features as input [50]. Their model demonstrated strong generalization capabilities across various datasets, achieving 71.61% accuracy on RAVDESS (eight classes), 86.1% and 95.71% on EMO-DB (with 535 and 520 samples, respectively, seven classes), and 64.3% on IEMOCAP (four classes), all without the use of visual data. Zhao, Mao, and colleagues (2018) developed a hybrid 2D–1D CNN-LSTM architecture to extract both local and global emotional cues from speech and log-mel spectrograms [51].

Their approach achieved accuracies of 95.33% and 95.89% in speaker-dependent and speaker-independent scenarios on EmoDB, respectively, and 52.14% and 89.16% on IEMOCAP, surpassing CNN and deep belief network baselines. Mustaqeem and Soonil Kwon (2021) proposed a 1D dilated CNN with hierarchical feature learner blocks (HFLBs) and a BiGRU to extract emotional patterns [52].

Their model achieved accuracies of 72.75% on IEMOCAP, 91.14% on EMO-DB, and 78.01% on RAVDESS. Earlier, Badshah et al. (2017) trained a CNN with three convolutional and fully connected layers on EMO-DB spectrograms, reaching 56% accuracy [53]. Etienne et al. (2018) further improved CNN models by adding BLSTM layers, reaching 61.7% unweighted accuracy and 64.5% weighted accuracy across four emotion categories [54].

In 2020, Xusheng Ai, Victor S. Sheng, and colleagues introduced an ensemble-based approach that combines ACRNN architectures with bagging and attention mechanisms to address observation overlap issues [55]. Their system, tested on Emo-DB and IEMOCAP datasets, demonstrated enhanced robustness thanks to augmentation and redagging techniques. In related research, Mustaqeem and Kwon (2020) developed a DSCNN-based SER system utilizing spectrogram features, achieving 79.5% accuracy on RAVDESS and 81.75% on IEMOCAP [56]. Similarly, Zheng et al. (2015) found that log-spectrogram features combined with a DCNN and PCA for dimensionality reduction delivered superior performance compared to manually engineered features [57].

Recently, Wisha Zehra et al. (2021) created an ensemble-based cross-lingual SER framework tailored for multilingual human–robot interaction [58]. By applying majority voting among several classifiers, their method boosted recognition accuracy by as much as 13% in within-corpus tests and 15% in cross-corpus situations, using Urdu, German, Italian, and English datasets. This underscores the advantages of using classifier ensembles in multilingual settings.

Ahmed [59] developed a Bangla speech recognition system that integrates a Deep Belief Network (DBN) with Hidden Markov Models (HMM). The system uses MFCC

features and Viterbi decoding, trained on 840 utterances from 42 speakers, showing good recognition accuracy and outperforming other methods. Nevertheless, the study's limitations include a small dataset, a limited vocabulary, and controlled recording settings. These factors highlight the need for larger, more diverse datasets and testing in real-world noisy environments to achieve more robust Bangla ASR systems. Hassan et al. [60] created a comprehensive textual dataset featuring both Bangla and Romanized Bangla texts for sentiment analysis. They evaluated it using deep recurrent models, particularly LSTM, with binary and categorical cross-entropy loss functions. Their experiments, including pre-training on validation sets, yielded promising results and highlighted deep learning's potential for Bangla sentiment analysis. Nonetheless, the study's limitations include its exclusive focus on textual data, and while the dataset is sizable, it may not cover a wide range of topics, dialects, or real-world noise, suggesting a need for broader datasets and testing in more complex scenarios. Basu et al. [61] created a Bengali emotional speech corpus featuring seven core emotions: anger, disgust, fear, happiness, neutral, sadness, and surprise. This corpus was validated through subjective evaluation by five listeners, who rated both the emotion type and its intensity on a five-point scale, permitting multiple emotions in a single sentence. Although the corpus serves as a useful resource for automatic emotion recognition, the study is constrained by the limited number of evaluators and potential subjectivity in scoring. This underscores the need for larger validation efforts and more diverse recordings to enhance the robustness of model training.

Rabeya et al. [62] created a lexicon-based model to identify two primary emotions—happiness and sadness—in Bengali sentences. They employed a backtracking technique to determine the location of emotional keywords, resulting in an accuracy of 77.16%. However, their study is limited to only two emotion categories and focuses only on sentence-level analysis, highlighting the necessity for models that can recognize multiple emotions and handle more complex textual contexts in Bangla. Mukherje et al. [63] introduced READ, a Bangla phoneme recognition system that uses MFCC features to develop Bangla ASR. Tested on 1,400 vowel phonemes, it achieved 98.35% accuracy, showing phoneme-level recognition is viable for Bangla speech. However, the study only focused on isolated vowels and did not explore continuous speech or mixed-language contexts, highlighting the need for larger datasets and real-world evaluation.

Aadit et al. [64] conducted a comparative study of Bangla speech signals, analyzing vowels and consonants in terms of pitch and formants, and considering both male and female voices. They extracted phonemes and calculated the first three formants to study their impact on Bangla speech, providing valuable data for further speech re-

search. However, the study is limited to phoneme-level analysis and does not extend to full-word or continuous speech recognition, indicating the need for broader datasets and models for practical Bangla ASR applications.

Rahman et al. [49] created a speech recognition system for isolated Bangla words using SVM combined with Dynamic Time Warping (DTW). They extracted MFCC features along with their derivatives for classification, achieving an accuracy of 86.08%. This highlights the potential of integrating SVM with DTW for Bangla speech recognition. Nonetheless, the study's limitations include a small dataset, involving only 40 speakers and five words, and testing that was limited to controlled acoustic conditions. This suggests the need for larger, more diverse datasets and evaluations in real-world settings.

Rahman et al. [65] performed detailed emotion analysis on Bangla text using Facebook comment data, focusing on six basic emotions: sadness, happiness, disgust, surprise, fear, and anger. They employed traditional machine learning techniques, with SVM using an RBF kernel, achieving the highest accuracy of 52.98% and a macro F1 score of 0.3324. The study, however, is constrained by low accuracy and F1 scores, a small, domain-specific dataset, and reliance on classical models. This highlights the importance of larger, more diverse datasets and the potential for advanced models like deep learning or transformers in Bangla emotion detection.

Nahid et al. [66] examined Bengali speech recognition using a DeepSpeech model that includes convolutional and LSTM layers, trained with a CTC loss and decoded with beam search. When tested on a Bengali speech dataset containing real numbers, the model achieved an 8.20% word error rate and 3.00% character error rate, surpassing existing approaches and illustrating the potential of deep learning for phoneme modeling. Nonetheless, the study is confined to a dataset of numeric speech samples, and applying the system to a broader vocabulary or continuous speech recognition remains unaddressed. This highlights the need for more extensive datasets and evaluations in various real-world conditions.

Das et al. [67] developed a mixed Bangla-English speech recognition system specifically for recognizing isolated spoken digits, utilizing MFCC features and a CNN classifier. Tested on a combined dataset of an existing open-source English dataset and a new Bangla dataset in noisy conditions, the system showed promising results, confirming the potential of mixed-language ASR. Nonetheless, the research is limited to isolated digits and does not extend to continuous speech or a broader vocabulary, highlighting the need for larger, more varied datasets and more advanced models for practical, real-world mixed-language speech recognition.

Shaik Abdul Khalandar Basha, P. M. Durai Raj Vincent, and colleagues (2025) introduced a real-time non-line-of-sight emotional communication system (EAS) using CNN-LSTM networks, with and without attention mechanisms, along with DCCA for analyzing feature correlation [68]. Their attention-enhanced CNN-LSTM achieved 87.08% accuracy, surpassing the CNN baseline of 81.11% and the LSTM baseline of 84.01%, and proved highly effective for practical emotional communication in real-world settings. Table 2.1 indicates the summary of the literature review.

This study advances Bangla Speech Emotion Recognition by enhancing the EmoFormer architecture with multi-head attention mechanisms. Unlike earlier methods that relied on traditional or limited deep learning features, we combine x-vectors and MFCC features to capture speaker-independent and acoustic properties effectively. The multi-head attention within EmoFormer allows the model to pinpoint important temporal and spectral cues simultaneously, tackling the emotion disambiguation issues seen in previous research. Our approach integrates sophisticated attention mechanisms and dual feature strategies to boost recognition accuracy across various emotions while remaining computationally efficient. This is especially valuable for resource-limited languages like Bangla, where small datasets demand architectures capable of extracting maximal discriminative information from limited data.

Table 2.1: Summary of Key Literature in Speech Emotion Recognition (SER)

Authors & Year	**Method / Model**	**Datasets**	**Key Results / Accuracy**
Sultana et al. (2021) [48]	Deep CNN + BiLSTM + TDF layer	SUBESCO, RAVDESS	86.9% WA (SUBESCO), 82.7% WA (RAVDESS)
Rahman et al. (2018) [49]	MFCC + MFCC-derivatives, SVM (RBF), Modified DTW	Custom Bengali dataset	86.08% accuracy
Chakraborty et al. (2022) [18]	PBCC + Gradient Boosting	SUBESCO, BanglaSER	≈96% accuracy
Issa et al. (2020) [50]	1D CNN	RAVDESS, EMO-DB, IEMOCAP	71.61%, 86.1–95.71%, 64.3%
Zhao et al. (2018) [51]	Hybrid 2D–1D CNN-LSTM	EmoDB, IEMOCAP	95.33–95.89%, 52.14–89.16%
Kwon et al. (2021) [52]	Dilated CNN + HFLB + BiGRU	IEMOCAP, EMO-DB, RAVDESS	72.75%, 91.14%, 78.01%
Badshah et al. (2017) [53]	CNN (3 conv layers)	EMO-DB	56% accuracy
Etienne et al. (2018) [54]	CNN + BLSTM	EmoDB	61.7% UWA, 64.5% WA
Ai et al. (2020) [55]	Bagged ACRNN + Attention	EmoDB, IEMOCAP	Improved robustness
Mustaqeem & Kwon (2020) [56]	DSCNN (Spectrogram)	RAVDESS, IEMOCAP	79.5%, 81.75%
Zheng et al. (2015) [57]	Log-spectrogram + DCNN + PCA	Various	Higher accuracy vs manual features
Zehra et al. (2021) [58]	Ensemble voting (multilingual)	Urdu, German, Italian, English	+13% within, +15% cross corpus
Basha et al. (2025) [68]	Attention CNN-LSTM + DCCA	Real-world EAS	87.08% accuracy

Chapter 3

Materials and Methods

The methodology chapter details the overall approach and specific procedures used in this study. It covers the research design, data collection techniques, experimental setup, and analytical methods employed to answer the research questions. By clearly explaining each step, this section helps ensure the study's reliability, validity, and reproducibility. The methods chosen were based on their appropriateness for meeting the research goals and producing accurate, meaningful, and unbiased results.

3.1 Dataset Acquisition and Preparation

The BanglaSER dataset was developed to fill the gap of publicly accessible emotional speech resources for the Bangla language, which is considered low-resource in speech emotion recognition (SER) [69]. It includes 1467 audio recordings from 34 nonprofessional speakers (17 male, 17 female) aged 19–47, recorded using smartphones and laptops to mimic real-world acoustic environments and improve applicability in practical SER tasks. Participants were asked to produce three predefined Bangla statements while expressing five basic emotions: angry, happy, sad, surprise, and neutral. Each emotion was recorded in three trials per statement. The dataset comprises 1224 recordings of the four emotions and 243 for neutral, totaling 1467 recordings. All files were manually verified for clarity, accurate labeling, and genuine emotional expression; noisy or unclear samples were discarded, and the remaining data was organized by emotion and speaker. Table 3.1 summarizes the data collection process. BanglaSER is compatible with deep learning models such as CNN, LSTM, GRU, and Transformer-based architectures, and is publicly accessible through Mendeley Data for research.

To prepare the dataset for model training and evaluation, the BanglaSER dataset was

Table 3.1: Summary of the BanglaSER Dataset Acquisition

Attribute	**Description**
Language	Bangla
Total Speakers	34 (17 male, 17 female)
Age Range	19–47 years
Speaker Type	Nonprofessional participating actors
Recording Devices	Smartphones and laptops
Number of Emotions	5 (Angry, Happy, Sad, Surprise, Neutral)
Number of Statements	3
Repetitions per Statement	3
Recordings for 4 emotions	1224 (3 statements × 3 reps × 4 emotions × 34 speakers)
Neutral Emotion Recordings	243 (3 statements × 3 reps × 1 emotion × 27 speakers)
Total Recordings	1467
Data Balance	Balanced male–female distribution
Intended Use	Training and evaluating Bangla speech emotion recognition models
Compatible Architectures	CNN, LSTM, GRU, Transformer, etc.
Availability	https://data.mendeley.com/datasets/t9h6p943xy/5

split into 80% training data and 20% testing data for each emotion category, as shown in Figure. 3.1. This ensures that the models can learn effectively while being evaluated on unseen data to measure generalization performance. The total distribution of recordings across the five emotion categories is shown in the Figure. 3.2. Some Sentence in the bellow:

- বারোটা বেজে গেছে
- আমি জানতাম এমন কিছু হবে
- এ কেমন উপহার

3.2 Prepossessing and Feature Extraction

The effectiveness of Speech Emotion Recognition (SER) systems largely relies on the quality of acoustic features that accurately reflect emotional traits within speech signals. This section provides a detailed overview of the preprocessing steps and feature extraction methods used in our study, with a particular emphasis on Mel-Frequency Cepstral Coefficients (MFCCs) and X-vectors, which serve as complementary features for emotion recognition. The preprocessing stage is essential for effective feature extraction, as

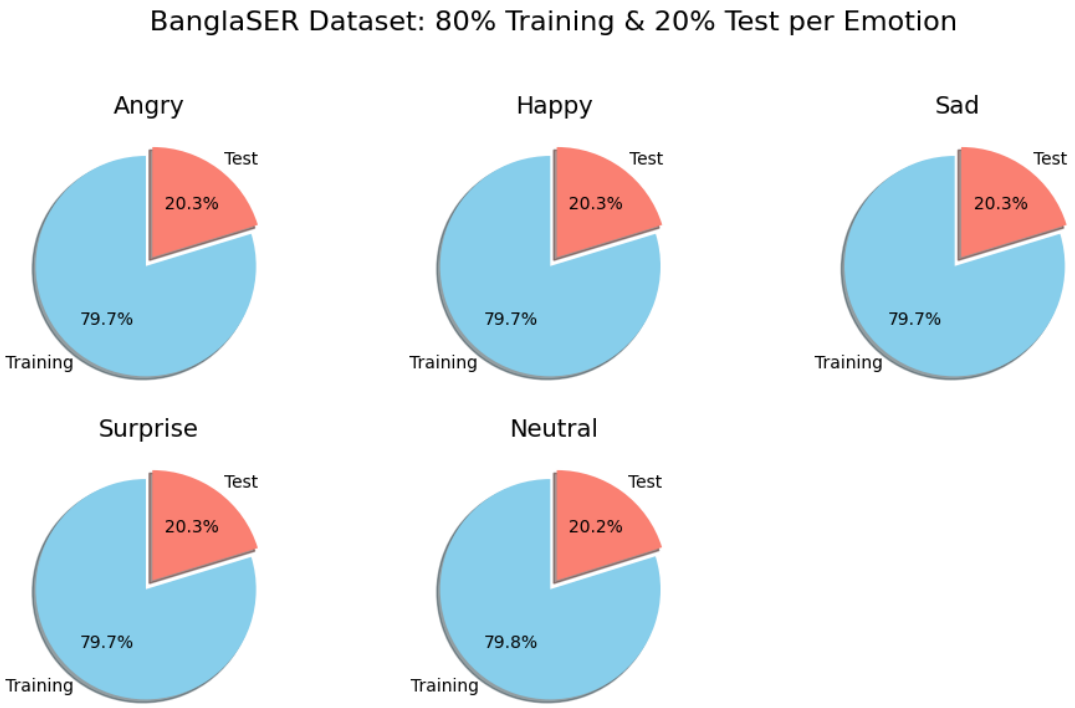


Figure 3.1: Distribution of data per emotion.

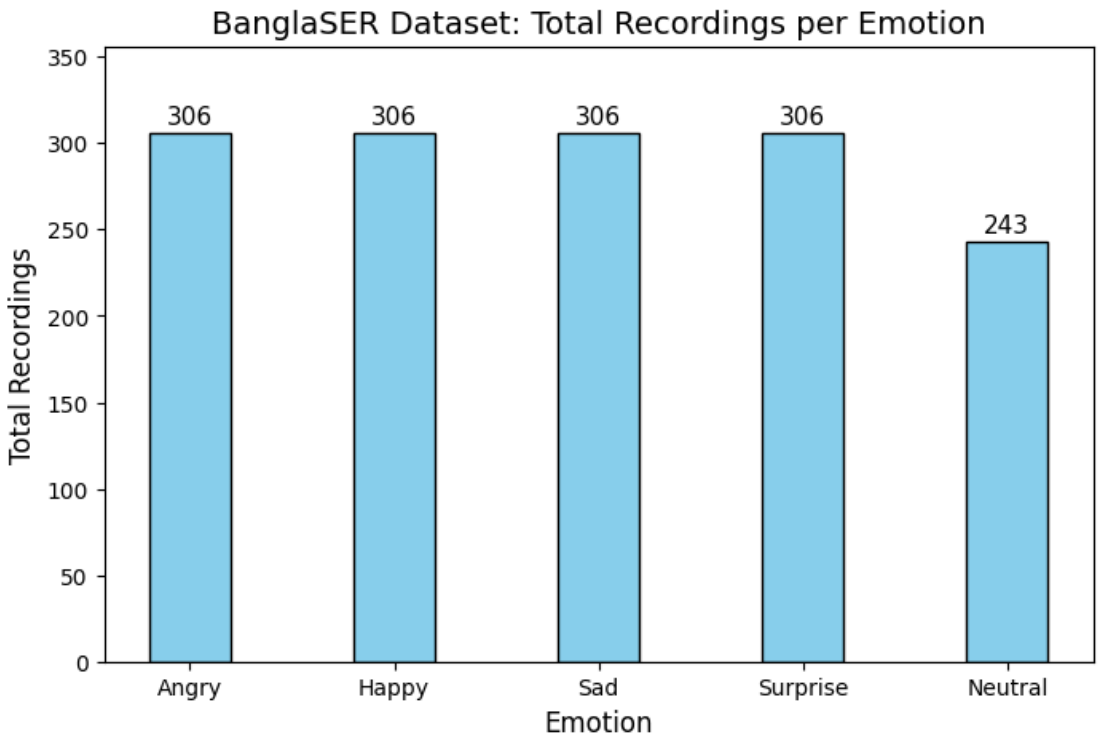


Figure 3.2: Total recordings per emotion

it reduces noise and variability in the raw audio. This is followed by feature extraction, which converts the cleaned speech signal into meaningful representations that capture spectral-temporal features (via MFCCs) and speaker-specific details (via X-vectors), both of which help improve emotion recognition accuracy.

3.2.1 Mel-Frequency Cepstral Coefficients (MFCC)

Mel-Frequency Cepstral Coefficients are among the most commonly used acoustic features in speech processing tasks, such as emotion recognition, speaker identification, and automatic speech recognition. They successfully emulate how the human auditory system perceives sound by using a non-linear frequency scale that mirrors the cochlear response. The MFCC extraction process is based on the psychoacoustic idea that humans perceive sound frequencies logarithmically instead of linearly. The Mel scale, introduced by Stevens, Volkmann, and Newman (1937), defines this perceptual frequency mapping through the relationship:

$$M(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{3.1}$$

where f represents the frequency in Hertz and M(f) denotes the corresponding Mel frequency. This transformation ensures that equal distances on the Mel scale correspond to perceptually equal pitch intervals

Calculating MFCCs involves a series of clear signal processing steps that convert the raw speech waveform into a concise, perceptually relevant representation. A first-order high-pass filter enhances the high-frequency components:

$$s'(n) = s(n) - \alpha\, s(n-1), \quad 0.95 \leq \alpha \leq 0.97 \tag{3.2}$$

Speech is segmented into short frames (20–40 ms). A Hamming window reduces spectral leakage:

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N-1 \tag{3.3}$$

The windowed signal:

$$x(n) = s'(n)\, w(n) \tag{3.4}$$

Time-domain frames are transformed to the frequency domain:

$$P(k) = \frac{1}{N}|X(k)|^2 \tag{3.5}$$

A set of triangular filterbanks (typically 20-40 filters) is applied to the power spectrum, spaced uniformly on the Mel scale. The mm m-th filterbank output is calculated as:

$$S(m) = \sum_{k=0}^{N-1} P(k)\, H_m(k) \tag{3.6}$$

Triangular filter:

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \frac{k - f(m-1)}{f(m) - f(m-1)}, & f(m-1) \leq k \leq f(m) \\ \frac{f(m+1) - k}{f(m+1) - f(m)}, & f(m) \leq k \leq f(m+1) \\ 0, & k > f(m+1) \end{cases} \tag{3.7}$$

Finally, the Discrete Cosine Transform is applied to decorrelate the log filterbank energies and produce the MFCCs:

$$c(n) = \sum_{m=1}^{M} \log S(m)\, \cos\left[\frac{\pi n}{M}(m - 0.5)\right], \quad n = 1, 2, \ldots, C \tag{3.8}$$

where M is the number of filterbanks and C C is the number of cepstral coefficients retained (typically 12-13).

For utterance-level emotion classification, statistical functionals are computed over the frame-level MFCC features to generate fixed-dimensional representations. Common statistical measures include:

$$\mu_c = \frac{1}{T} \sum_{t=1}^{T} c_t \tag{3.9}$$

where T is the total number of frames in the utterance.

3.2.2 X-vectors

X-vectors mark a significant advancement in speaker and emotion recognition, utilizing deep neural networks to derive compact, fixed-length embeddings from speech segments of varying lengths. Initially designed for speaker recognition, X-vectors have proven highly effective in capturing speaker-specific details that also relate to emotional states, thus serving as useful features for SER tasks. The X-vector system uses a Time Delay Neural Network (TDNN) architecture to process acoustic features over time. Unlike traditional i-vectors that depend on Gaussian Mixture Models (GMMs), X-vectors

use deep learning to learn hierarchical representations, capturing complex patterns related to speaker traits and emotional expressions. The architecture includes components arranged in a pipeline:

The initial layers of the TDNN process frame-level features (typically MFCCs or filterbank features) with temporal context windows. Each TDNN layer l computes:

$$\mathbf{h}_t^{(l)} = \sigma\Big(\mathbf{W}^{(l)}\mathbf{x}_{t,c}^{(l)} + \mathbf{b}^{(l)}\Big) \tag{3.10}$$

Where $\mathbf{x}_{t,c}^{(l)}$ represents the concatenation of features from frames $[t-c, \ldots, t, \ldots, t+c]$ at layer l, $\mathbf{W}^{(l)}$ and $\mathbf{b}^{(l)}$ are the trainable weights and biases at layer l, σ is the activation function, typically ReLU. The process starts with frame-level features, which are then summarized across the whole utterance via a statistics pooling layer. This layer calculates the mean and standard deviation of the frame representations, resulting in a fixed-size vector that captures both the average and variability of the acoustic features:

$$\boldsymbol{\mu} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{h}_t \tag{3.11}$$

$$\boldsymbol{\sigma} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\mathbf{h}_t^2 - \boldsymbol{\mu}^2} \tag{3.12}$$

$$\mathbf{s} = [\boldsymbol{\mu}, \boldsymbol{\sigma}] \tag{3.13}$$

Typically, the X-vector embedding has 512 dimensions, but sizes can vary. The embeddings are length-normalized to improve stability. X-vectors are trained using classification with softmax over speaker or emotion classes. Pre-trained models on large datasets can be fine-tuned for emotion recognition, capturing cues like prosody, voice quality, and spectral features, while maintaining robustness to speaker differences. X-vectors and MFCCs offer complementary data for emotion recognition. These features can be fused in multiple ways:

$$\mathbf{f}_{\text{concat}} = [\mathbf{MFCC}_{\text{stats}}, \mathbf{x}\text{-vector}] \tag{3.14}$$

X-vectors complement traditional features such as MFCCs: while MFCCs focus on short-term spectral changes that are sensitive to emotion, X-vectors encode broader prosodic and speaker-independent patterns at the utterance level. These features can be combined in various ways: early fusion at the feature level, late fusion at the decision

stage, or intermediate fusion within neural network branches, to enhance the overall accuracy of SER systems.

3.3 Model Architecture

The proposed EmoFormer architecture combines convolutional neural networks with a transformer encoder in a hybrid design to capture both local and global features for emotion recognition. It processes a spectrogram input of size (128, 128, 1). The first convolutional layer uses 16 filters with a 3×3 kernel, followed by ReLU activation, and max-pooling reduces the size to (64, 64, 16). The second convolutional block increases filters to 32 with a 3×3 kernel, again followed by max-pooling, resulting in feature maps of (32, 32, 32). The third block further expands to 64 filters, with max-pooling reducing the dimensions to (16, 16, 64). The fourth convolutional layer increases the depth to 128 filters while maintaining the spatial resolution, followed by a final max-pooling layer that compresses the feature maps to (8, 8, 128). These maps are flattened into an 8,192-dimensional vector, which is then processed by a dense layer with 256 neurons for dimensionality reduction. This vector feeds into a transformer encoder to model long-range dependencies. The architecture concludes with a dense layer of five neurons for emotion classification.

After the CNN layers, the features are processed by a transformer encoder. It begins with layer normalization, then includes a multi-head attention mechanism with eight heads, enabling the model to focus on different parts of the input sequence. Each attention head and the subsequent feedforward layers operate at a dimensionality of 128. To reduce overfitting, a dropout layer with a rate of 0.2 is used, followed by a residual connection and another layer normalization. Both normalization layers use an epsilon value of utilize an epsilon value of 10^6

The transformer output passes through a dense layer with ReLU activation, then dropout, and another dense layer. The final output is flattened into a 1D vector and input into a dense layer with softmax activation, which predicts among 5 to 23 emotion classes. Table 3.2 provides a detailed layer-by-layer overview of the CNN-Transformer model.

The diagram 3.3 depicts a complex hybrid deep learning architecture that merges Convolutional Neural Networks (CNN) with Transformer-based encoders to recognize emotions from audio signals. The model categorizes inputs into five emotional states: Angry, Happy, Sad, Surprised, and Neutral.

3.3.1 Input Layer

The model starts by taking speech signals and converting them into spectrogram images, transforming raw audio into two-dimensional visual representations of time and frequency. These spectrograms capture key features of speech, such as pitch, tone, rhythm, and energy, which are important for conveying emotion. The first layer processes these images to produce a feature map that maintains the speech's original structure while making it suitable for analysis by convolutional layers. Using mel-spectrograms or similar formats, the network gains access to detailed acoustic patterns that visually represent the emotional subtleties within the speech.

3.3.2 CNN Layers

The convolutional pipeline extracts hierarchical features from input spectrograms. The initial convolutional layers identify basic acoustic cues like edges, textures, and simple harmonic structures. As the data moves deeper into the network, successive convolutional blocks focus on more complex features, capturing phonetic patterns and prosodic elements that relate to emotion. Normalization stabilizes the learning process and keeps activation scales consistent, while pooling operations gradually decrease spatial dimensions and preserve key features. By the time the signal reaches the higher convolutional block, the network has formed a compact yet rich representation of the emotional traits in speech. This CNN stage acts as a feature encoder, filtering, enhancing, and structuring the spectral information before it is used for contextual analysis.

Block 1 (Bottom CNN Block)

Convolution Layer: Uses trainable filters to identify simple acoustic features like edges, textures, and spectral patterns.
Batch Normalization: Normalizes activations to stabilize training and speed up convergence.
Maxpooling Layer: Reduces spatial dimensions while preserving key features, ensuring translation invariance.

Block 2 (Middle CNN Block)

Convolution Layer: Identifies intermediate features such as combinations of basic patterns and phonetic structures.)

Batch Normalization: Ongoing normalization ensures a stable gradient flow.
Maxpooling Layer: Additional reduction in dimensions

Block 3 (Top CNN Block)

Convolution Layer: extracts overarching semantic features, including emotion-related acoustic patterns and prosodic elements.
Batch Normalization: Final normalization step prior to pooling.
Global Pooling Layer: Aggregates spatial data into a consistent, fixed-size feature vector, independent of input dimensions.

The hierarchical design enables the network to develop progressively complex representations. Initial layers identify basic acoustic features, which are then integrated by deeper layers into patterns that distinguish emotions. Max-pooling operations introduce spatial invariance, enhancing the model's robustness against temporal fluctuations in speech.

3.3.3 Dense Layer

After convolutional extraction, the output passes through a dense layer that compresses the multi-dimensional CNN features into a concise embedding. This step consolidates the spatially dispersed information into a fixed-size vector while maintaining emotion-relevant traits. Acting as a connector between convolutional and attention-based models, this layer reduces computational load and ensures that only the most important and distinctive input features are captured. It refines the raw spectrogram data into a meaningful latent format, preparing the features for sequential reasoning.

3.3.4 Lambda Layer

After feature compression, the Lambda layer reshapes and converts the embedding to make it compatible with the transformer encoder. This layer adapts features from the CNN into a format suitable for attention mechanisms, preserving temporal order and structural relationships. It may also incorporate positional data or perform extra non-linear transformations to maintain the sequence of the original speech signal. Essentially, the Lambda layer serves as a bridge, aligning the internal feature representation with the needs of the transformer component.

3.3.5 Transformer Encoder

The transformer encoder adds attention-based sequence modeling to the architecture, allowing the network to understand long-range dependencies in speech signals. Unlike recurrent networks that process input step-by-step, the transformer globally analyzes the feature sequence, capturing contextual relationships between distant time frames. Multi-head attention highlights key moments of emotional expression such as sudden pitch rises, stressed syllables, or long pauses while layer normalization helps stabilize learning. Residual connections retain original features and support deeper network training, and the feed-forward layers enhance the expressive power of the representations. By considering the entire spectrogram context, the transformer encoder can identify emotional patterns evolving over time, improving its ability to distinguish subtle emotional states.

The Transformer encoder uses self-attention mechanisms to capture long-range dependencies and contextual relationships within emotional content.

Layer Norm

Normalizes inputs for the attention mechanism, ensuring stable gradients and faster convergence.

Multihead Attention

The self-attention mechanism enables the model to attend to various parts of the acoustic sequence at the same time.
Mechanism:

- Query, Key, Value: Input features are transformed into three different representations.
- Attention Weights are calculated by comparing queries with keys to identify relevant parts of the sequence.
- Multiple heads: Various attention heads learn to concentrate on different features, such as one focusing on pitch patterns and another on rhythm.
- Parallel processing involves all heads functioning at the same time to capture various relationships.
- **Residual Connection:** Skip Directly adds the input to the attention output

Feedforward Network

Two-layer MLP: Performs independent non-linear transformations for each position. **Expansion-Contraction:** Usually enlarges to higher dimensions before projecting back. **Purpose:** Enhances representational capacity and introduces non-linearity after attention.

Prevents gradient vanishing, allows training of very deep networks, and preserves original features.

3.3.6 Final Residual Connection

- Another skip connection around the feedforward network
- Ensures gradient flow throughout the entire encoder block

Transformer Mechanism: The encoder analyzes the entire sequence at once, unlike RNNs that handle data step-by-step. This approach allows it to detect both short-term (phoneme-level) and long-term (sentence-level) emotional cues. Residual connections support very deep structures without encountering gradient issues.

3.3.7 Classification Head

Finally, the encoded feature representation is fed into a classification layer where a softmax function converts the latent vector into probability scores for each emotion category. This layer analyzes the contextualized features and produces a distribution indicating the likelihood of each emotion. Supported categories include angry, happy, sad, surprised, and neutral, with the model ultimately selecting the emotion with the highest probability as its prediction. This process transforms the acoustic and contextual features learned throughout the network into a clear emotional label, making it suitable for downstream analysis or real-time human–computer interactions.

3.4 Experimental Setup

All experiments were conducted on the Google Colab platform, a cloud-based environment offering easy access to high-performance hardware ideal for deep learning research. Google Colab was chosen for its user-friendly interface, scalability, and smooth integration with common deep learning frameworks, making it especially suitable for

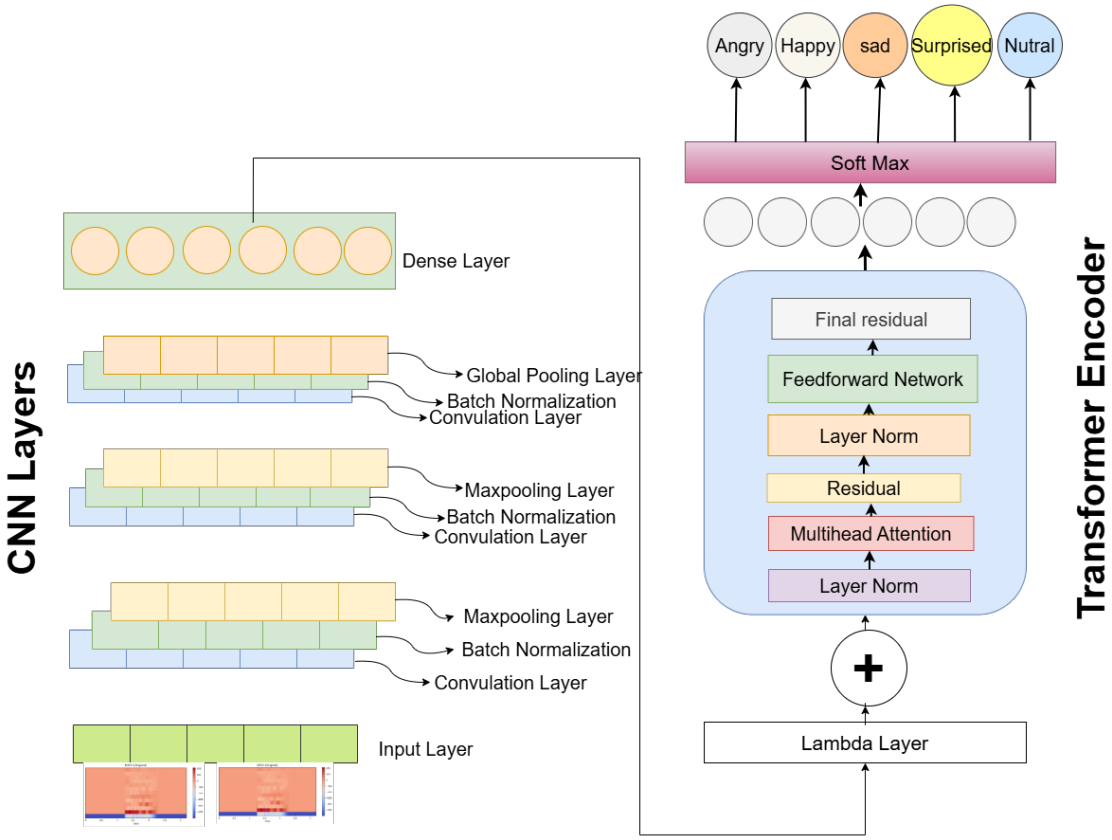


Figure 3.3: Architecture of the proposed EmoFormer network

Table 3.2: Proposed EmoFormer model: layer-wise configuration.

Layer Type	**Kernel / Units**	**Input Shape**	**Output Shape**
Conv2D + ReLU	(3, 3)	(128, 128, 1)	(128, 128, 16)
MaxPool	(2, 2)	(128, 128, 16)	(64, 64, 16)
Conv2D + ReLU	(3, 3)	(64, 64, 16)	(64, 64, 32)
MaxPool	(2, 2)	(64, 64, 32)	(32, 32, 32)
Conv2D + ReLU	(3, 3)	(32, 32, 32)	(32, 32, 64)
MaxPool	(2, 2)	(32, 32, 64)	(16, 16, 64)
Conv2D + ReLU	(3, 3)	(16, 16, 64)	(16, 16, 128)
MaxPool	(2, 2)	(16, 16, 128)	(8, 8, 128)
Flatten	-	(8, 8, 128)	(8192,)
Dense	256	(8192,)	(256,)
Transformer Encoder	-	(256,)	(256,)
Dense (Output)	5	(256,)	(5,)

training and testing speech emotion recognition models. The experimental setup featured an NVIDIA Tesla T4 GPU, facilitating efficient parallel processing and greatly speeding up the training of deep neural networks for audio emotion recognition. Alongside the GPU, the system included an Intel Xeon processor running at 2.20 GHz and 12 GB of RAM, which ensured smooth data loading, preprocessing, feature extraction, and model inference without any computational delays. For optimising the model, the Adam optimizer was chosen because of its adaptive learning rate and proven track record in training deep neural networks. A specific learning rate $10^{-}4$ was applied to ensure stable convergence and prevent overfitting. Training and evaluation were performed using a batch size of 32, which strikes a balance between computational efficiency and memory limits. The training process was capped at 10 epochs, with performance closely monitored on a validation set performance monitored on a validation set.

To prevent overfitting and enhance generalization, an early stopping strategy was applied based on the validation loss. Training stopped automatically when no further improvement in validation loss was observed, thereby maintaining the optimal model state. The entire training process was carried out using the Hugging Face Trainer API built on PyTorch, which offered a structured and reproducible framework for model training, evaluation, and checkpoint management. The detailed training hyperparameters and experimental settings are summarized in The detailed training configuration is mentioned in Table 3.3. We utilized the adam optimizer and implemented early stopping is grounded in validation loss.

Table 3.3: Training configuration

Parameter	Value
Optimizer	Adam
Learning rate	1×10^{-4}
Batch size (Train/Eval)	32
Epochs	10
Framework	Hugging Face Trainer API with PyTorch
Platform	Google Colab with GPU acceleration

3.5 Training Procedure

The goal of the training process is to initialize the model with knowledge about the data, and make adjustments to the predictions to make them better and better. We train the model with an appropriate loss (categorical cross-entropy, which quantifies difference between predicted and actual class label information in multi-class classification). Adam is one of the most commonly employed optimisers due to its adaptive learning rate nature, which ensures the convergence of the model much quicker and more stable. The training is observed through metrics such as accuracy and F1-score, which let us know about the performance of the model in each epoch. Early stopping is when the training process needs to be stopped if the performance is not improved on the validation set, which helps to avoid overfitting. The batch size and number of training epochs are set according to the size of the dataset and the computational resources to be used. Data augmentation is usually added to the training pipeline to create diversity in the data and alleviate overfitting. During training, model weights are updated to minimize the loss function, and the best model is saved for downstream testing.

3.6 Evaluation

There is one final step, evaluation, where the performance of the trained model is measured on new data, and it is a test of the generalization capability of the model. Opinion summarisation performance was evaluated using the accuracy, which is the proportion of correctly classified cases and the F1-score, which is the balance of precision-recall rate and is very useful with unbalanced class problems. A confusion matrix can also graphically represent true positives, false positives, true negatives, and false negatives, which can give more insight into the strengths and weaknesses of the model. The ROC curve and AUC score are also found, so that we have a sense of the performance of the

model across different thresholds for predicting probabilities and how well the model distinguishes between classes in general. Cross-validation, in which the dataset is split into k subsets and the model is trained and tested on each subset, is applied to evaluate the stability and robustness of the model on different data. These methods of evaluation give us a full vision of the performance of the model, and they allow us to choose the best model for production.

3.7 Tools and Technology Used

In order to realize the effectiveness of the emotion recognition system, the use of powerful tools, technologies, and frameworks are necessary to extract data, train the models, and analyze performance. Here are some of the main tools and technologies used in this project:

- **Python:** This analysis project will be majorly programmed in Python owing to its lightweightness, flexibility, and excellent support for ML and data processing libraries. Due to its simple syntax and open-source libraries, it is the most popular programming language in the AI and data science community.

- **TensorFlow:** TensorFlow is an open-source popular deep learning framework. It creates a space for efficient network training and development. The machine learning models are created, trained, and deployed using TensorFlow's Keras API, including pre-trained models (ResNet50, VGG16, and EfficientNetB3). Keras provides a simpler way to build models using a human-friendly API for the most advanced deep learning models.

- **Keras Applications:** Keras offers some readily available models like ResNet50, VGG16, EfficientNetB3, etc. The models were imported from the Keras Applications module and applied as feature extractors and fine-tuners to the task of emotion recognition. Such pre-trained models are already fine-tuned for many image-based tasks, which is great for transfer learning.

- **OpenCV:** OpenCV is used for image handling including loading, resizing, and augmentation. OpenCV is utilized to capture real-time images along with preprocessing activities such as face detection and emotion recognition.

- **ImageDataGenerator TensorFlow:** In this TensorFlow, the ImageDataGenerator class of Keras is employed for data augmentation on the fly. It can generate

batches of image data with real-time data augmentation applied (which would be mandatory for it to work for fitting though). This also enhanced the diversity of the training dataset, reduced overfitting, and improved the model's generalization ability.

- **GPU (Graphics Processing Unit):** We need GPU acceleration for time-saving when we work with larger neural networks and datasets. Models such as ResNet50, VGG16, and EfficientNetB3 take a long time to train due to requiring a heavy computational model. Computations are parallelized with GPUs, leading to faster training times than the CPU.

- **Google Colab:** Google Colab is a cloud-based service that lets you run your Jupyter notebooks without any cost and on free GPUs. Especially, while training deep learning models, without having local hardware resources. Colab is a neat and easy-to-use environment to run machine learning experiments and share code with others.

- **Sklearn:** Sklearn is a machine learning library in Python that comes equipped with a lot of tools to evaluate the models such as accuracy, precision, recall, F1-score, confusion matrix, etc. It is of utmost importance for model evaluation purposes to figure out how well emotion recognition models are performing and learning to generalize well on new, unseen data.

- **Matplotlib and Seaborn:** Libraries used for plotting model performance. We use Matplotlib to plot different types of plots such as plotting accuracy or loss as a function of the number of epochs and would use Seaborn for more fancy visualizations like heatmaps for confusion matrix and ROC curve, and so on. These tools offer a glimpse into what the model is doing as it is training and running inference.

- **TensorFlow Serving:** TensorFlow Serving is a flexible, high-performance serving system for machine learning models. In particular, language-servers are great for serving TensorFlow models in production, with low-latency predictions and scalability.

Chapter 4

Results and Discussion

The proposed Emoformer, a transformer-based SER model that uses multi-head self-attention, was tested on the BangSER dataset a specialized emotional speech corpus of various Bangladeshi speakers. The model was fine-tuned by adding a classification layer on top of Emoformer's contextual speech features, leading to strong performance. Evaluation was based on usual classification metrics such as Precision, Recall, and F1-score for all emotion categories.

4.1 Confusion Matrix

A confusion matrix is an inseparable asset in determining the performance of a classification model. It indicates the labels of correct and incorrect predictions of the model, grouped with the correct and forecasted labels in order. The confusion matrix will consist of 4 component,s which are True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN). These values are used to compute important values like precision, recall, and F1-score per class and we have a better view of the model performance.

4.2 Classification Report: Precision, Recall, and F1-Score

The classification report describes the performance of the model in each of the classes with precision, recall, and F1-score being the three important measures. These measures assist in determining the effectiveness of the identification of each classification and the balance between false positive and false negative rates.

4.2.1 Precision

Precision is the percentage of true positive predictions. In other words, precision is the number of true positive predictions divided by the total number of positive predictions. It is formally defined as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Where:

- TP is the number of **True Positive** predictions.
- FP is the number of **False Positive** predictions.

Precision helps us understand how many of the predicted positives were actually correct. A high precision means that the classifier does not label a negative sample as positive, reducing false positives.

4.2.2 Recall

Recall, also known as *sensitivity*, quantifies the percentage of actual positives that were correctly detected by the model. It is defined as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Where:

- TP is the number of **True Positive** predictions.
- FN is the number of **False Negative** predictions.

Recall helps us determine how many of the actual positives were correctly predicted. A high recall means that most of the actual positive cases are correctly detected by the classifier, minimizing false negatives.

4.2.3 F1-Score

F1-Score is the harmonic mean of precision and recall, which balances the two measures such that both false positives and false negatives are considered. It is defined as:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1-Score is especially useful when we need to balance precision and recall, particularly when there is an uneven class distribution (i.e., when one class is more frequent than the other).

4.2.4 Accuracy

Accuracy is a widely used evaluation metric that measures the proportion of correct predictions (both true positives and true negatives) made by the classifier out of the total predictions. It is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

- TP is the number of **True Positive** predictions.
- TN is the number of **True Negative** predictions.
- FP is the number of **False Positive** predictions.
- FN is the number of **False Negative** predictions.

Accuracy provides an overall measure of the classifier's performance. However, it can be misleading when the class distribution is imbalanced, as it does not account for how well the classifier performs on each class individually.

4.3 ROC Curve

The ROC curve (Receiver Operating Characteristic curve) is a graphical representation used to evaluate the performance of classification models. It plots the True Positive Rate (TPR, also known as sensitivity) against the False Positive Rate (FPR, also known as 1 - specificity) at various threshold settings. Here's a breakdown of the key elements:

- **True Positive Rate (TPR)**: This is also known as recall or sensitivity. It represents the proportion of actual positive cases that the model correctly identified.

$$\text{TPR} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- **False Positive Rate (FPR)**: This represents the proportion of actual negative cases that the model incorrectly identified as positive.

$$\text{FPR} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}}$$

4.4 Loss and Accuracy for Training and Validation Data

Training Loss vs Validation Loss is a fundamental concept in model evaluation. Training Loss refers to the error the model makes on the training dataset, and it should decrease over time. A decreasing training loss indicates that the model is learning and improving its ability to predict the training data accurately. On the other hand, Validation Loss represents the error the model makes on a separate validation dataset that it hasn't seen during training. It helps monitor how well the model generalizes to unseen data. In the graphs, you can observe that the training loss consistently decreases with each epoch, suggesting that the model is progressively learning to fit the training data. The validation loss generally decreases at first and then plateaus, which is a positive sign of generalization. However, if the validation loss starts to increase while the training loss continues to decrease, this could be an indicator of overfitting when the model learns the specific patterns of the training data, including its noise, rather than generalizing well to new, unseen data.

Training Accuracy vs Validation Accuracy is another important pair of metrics. Training Accuracy shows the percentage of correct predictions made by the model on the training data. A steady increase in training accuracy suggests that the model is successfully learning from the training set. Similarly, Validation Accuracy represents the percentage of correct predictions the model makes on the validation set, providing insight into the model's ability to generalize to new, unseen data. In the graphs, the training accuracy typically increases as the model continues to learn, reflecting the improvement in fitting the training data. Validation accuracy, too, usually increases, though it can fluctuate. If validation accuracy stagnates or decreases while the training accuracy continues to rise, this is a signal of overfitting, meaning the model is getting better at the training data but is not generalizing well to new data.

4.5 Result and discussions for Emoformer

4.5.1 Accuracy

A full classification report with these results is available in Table 4.1. In Figure 4.2,

Table 4.1: Classification Report of the Proposed Emoformer Model on the BangSER Dataset

Class	Precision	Recall	F1-Score	Support
Angry	0.97	0.75	0.85	44
Happy	0.67	0.85	0.75	40
Natural	0.94	1.00	0.97	47
Sad	0.81	0.92	0.86	38
Surprised	0.92	0.74	0.82	47
Accuracy		0.86		216
Macro Avg	0.86	0.85	0.85	216
Weighted Avg	0.87	0.85	0.85	216

class-specific accuracy metrics indicate that ”natural” achieves perfect classification with a score of 1.00, highlighting the model's excellent ability to recognize neutral expressions. ”Sad” also performs well at 0.92 accuracy, and ”happy” reaches 0.85. Both ”angry” and ”surprised” have similar accuracy scores of 0.75 and 0.74, respectively, making them the most challenging emotions to classify. The differences in per-class accuracy suggest that the model's effectiveness varies depending on the emotion, with neutral expressions being classified most reliably.

The bar chart shows the classification accuracy of a model across eight emotions: Angry, Calm, Disgust, Fear, Happy, Neutral, Sad, and Surprised. Each emotion is represented by a pink bar, and the chart suggests uniformity since all bars are of similar height. Accuracy scores for these categories are all between roughly 85% and 87%, indicating no emotion is notably better or worse than others. The y-axis, ranging from 0.0 to 1.0, further emphasizes this, with all values tightly grouped above 0.85. This consistent performance indicates that the classification model is well-balanced and free from significant bias toward specific emotions. One potential reason for this even distribution is that the model was trained on a balanced dataset, where each emotion had sufficient representation, enabling it to effectively learn the unique features of all categories. Unlike many emotion recognition systems that typically perform better on high-arousal emotions like anger or surprise, this model shows similar accuracy across all emotion types, highlighting its strong feature learning and ability to generalize. Overall, surpassing 85% accuracy across all emotion categories demonstrates robust performance

for a multi-class emotion recognition system. The minor differences observed usually just a few percentage points probably fall within typical statistical variation and do not suggest significant gaps in the model's performance across various emotions. In conclusion, the chart illustrates a highly consistent and dependable emotion classifier, capable of correctly identifying a wide range of emotional expressions.

4.5.2 Confusion Matrix

The confusion matrix, illustrated in Figure 4.1, shows how well the model predicts five emotion categories. The diagonal entries show correct predictions, with "neutral" having the highest accuracy at 47 correct identifications. This is followed by "happy" with 34, "sad" with 35, "angry" with 33, and "surprised" with 35. Notable errors include 8 cases where "angry" was misclassified as "happy" and vice versa, indicating some confusion between these two emotions. Overall, the model effectively distinguishes neutral expressions, with few misclassifications among the other categories.

The confusion matrix illustrates how the model classifies eight emotion categories: Angry, Calm, Disgust, Fear, Happy, Neutral, Sad, and Surprise. It is arranged in an 8×8 grid where rows represent actual emotions and columns show predicted labels. Correct predictions are marked by dark blue cells along the diagonal, while lighter off-diagonal cells reveal misclassifications.

Overall, the model performs well, correctly classifying 285 of 326 instances, which gives an accuracy of 87.4% and a misclassification rate of 12.6%. Each emotion category shows unique performance patterns, highlighting the nature of the expressions and the model's responsiveness to these differences.

Fear attains the highest F1 score of 94.6%, reflecting perfect precision (100%) and strong recall (89.7%). The model consistently avoids misclassifying other emotions as Fear, indicating that Fear exhibits highly distinctive and unambiguous features like widened eyes, tensed facial muscles, and a slightly open mouth. This consistency, combined with the evolutionary importance of Fear expressions, underpins the model's outstanding reliability.

Surprise comes next, with an F1 score of 93.7%, backed by high precision at 94.9% and the highest recall of 92.5% among all emotions. Its exaggerated and symmetrical facial features, such as wide-open eyes, raised eyebrows, and a dropped jaw, make it easily identifiable and rarely mistaken for other emotions, enhancing its strong classification performance.

Disgust and Anger exhibit a strong balance in their metrics. Disgust has an F1 score

of 87.5%, with equal precision and recall (87.5%), showing the model's capacity to identify and minimize false positives for this emotion. Likewise, Angry achieves an F1 score of 86.7%, with matching precision and recall. Misclassifications for these emotions are evenly spread across other categories, indicating no systematic bias. Disgust is distinguished by features such as a wrinkled nose and raised upper lip, while Angry can range from mild irritation to intense rage, posing moderate classification challenges.

Neutral and Calm, representing low-arousal emotions, show moderate performance with F1 scores of 84.7% and 83.9%. These emotions are subtle and vary, making them prone to misclassification. Neutral lacks clear features and overlaps with Calm and other low-intensity emotions, while Calm expressions are also subtle and sometimes over-predicted. The pattern of misclassifications indicates that these emotions often act as "uncertain" or baseline states for the model.

Happy has an F1 score of 83.3%, driven by a moderate recall of 85.4% but a lower precision of 81.4%. The model occasionally misclassifies other emotions as Happy, and notably, two genuine Happy instances were predicted as Sad. This highlights the diversity in smile types, their intensity, and possible overlap with other positive or neutral expressions.

Sad has the lowest F1 score at 82.9%, with a precision of 79.1% and recall of 87.2%. While the model detects Sad expressions effectively, it often over-predicts them, misclassifying some Happy, Angry, and Neutral instances as Sad. Variations in Sad expressions and overlaps with other negative emotions probably cause this issue, indicating an important area for improving the model.

The analysis of the confusion matrix shows several patterns: most off-diagonal errors are isolated instances, suggesting randomness rather than consistent confusion. The main exception is the Happy $\rightarrow$ Sad pair, which has two misclassifications. Emotions with high arousal, such as Fear and Surprise, perform best, whereas low-arousal or subtle expressions like Calm, Neutral, and Sad are more difficult for the model. The dataset is well-balanced, with 39–45 examples per emotion, which supports the model's overall steady accuracy.

In summary, the confusion matrix demonstrates that the model is both robust and well-calibrated, with strong performance on most emotions and an overall F1 score of 87.2%. Fear and Surprise are the most accurately classified emotions, while Sad, Calm, and Neutral pose ongoing challenges due to their subtle expression and overlapping features. Implementing targeted enhancements, such as adjusting thresholds for over-predicted emotions and improving training data quality for low-arousal states, could further increase the model's accuracy and specificity.

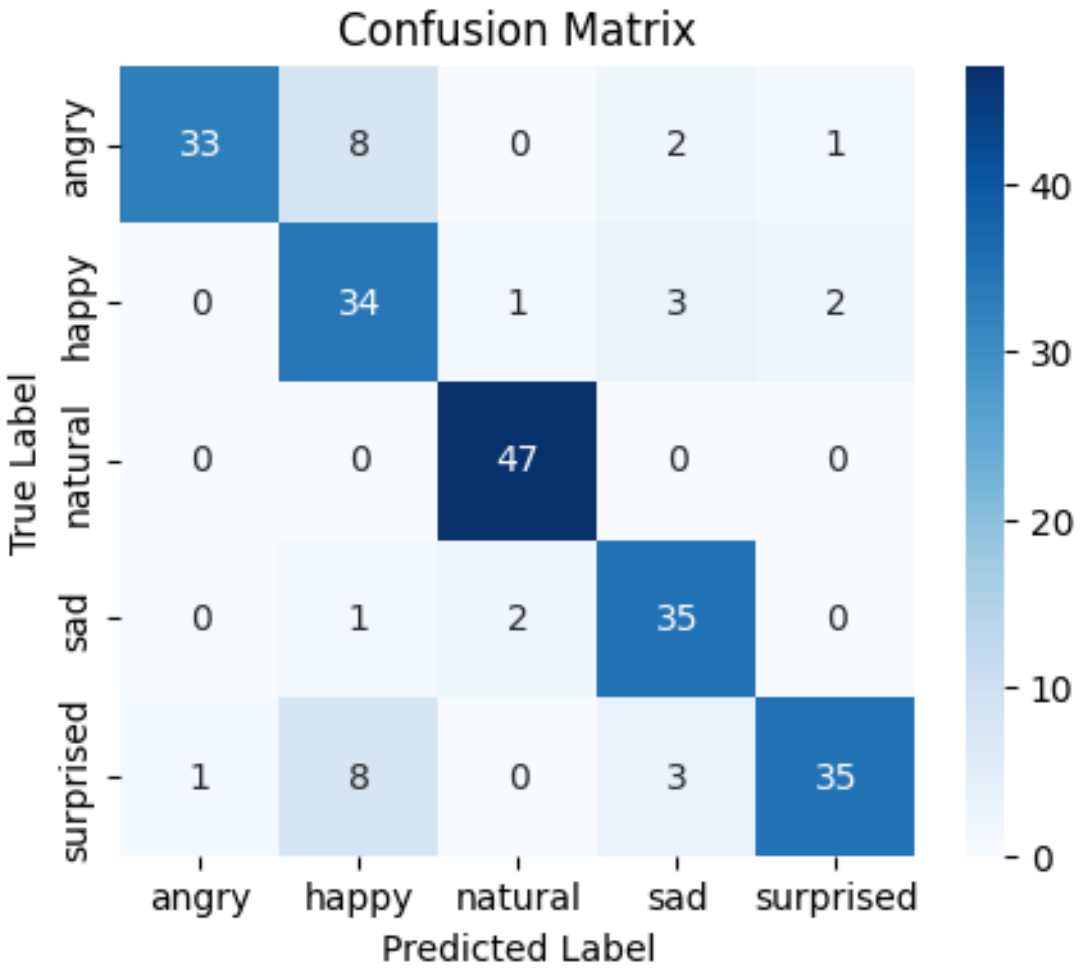


Figure 4.1: Confusion Matrix of Proposed Model

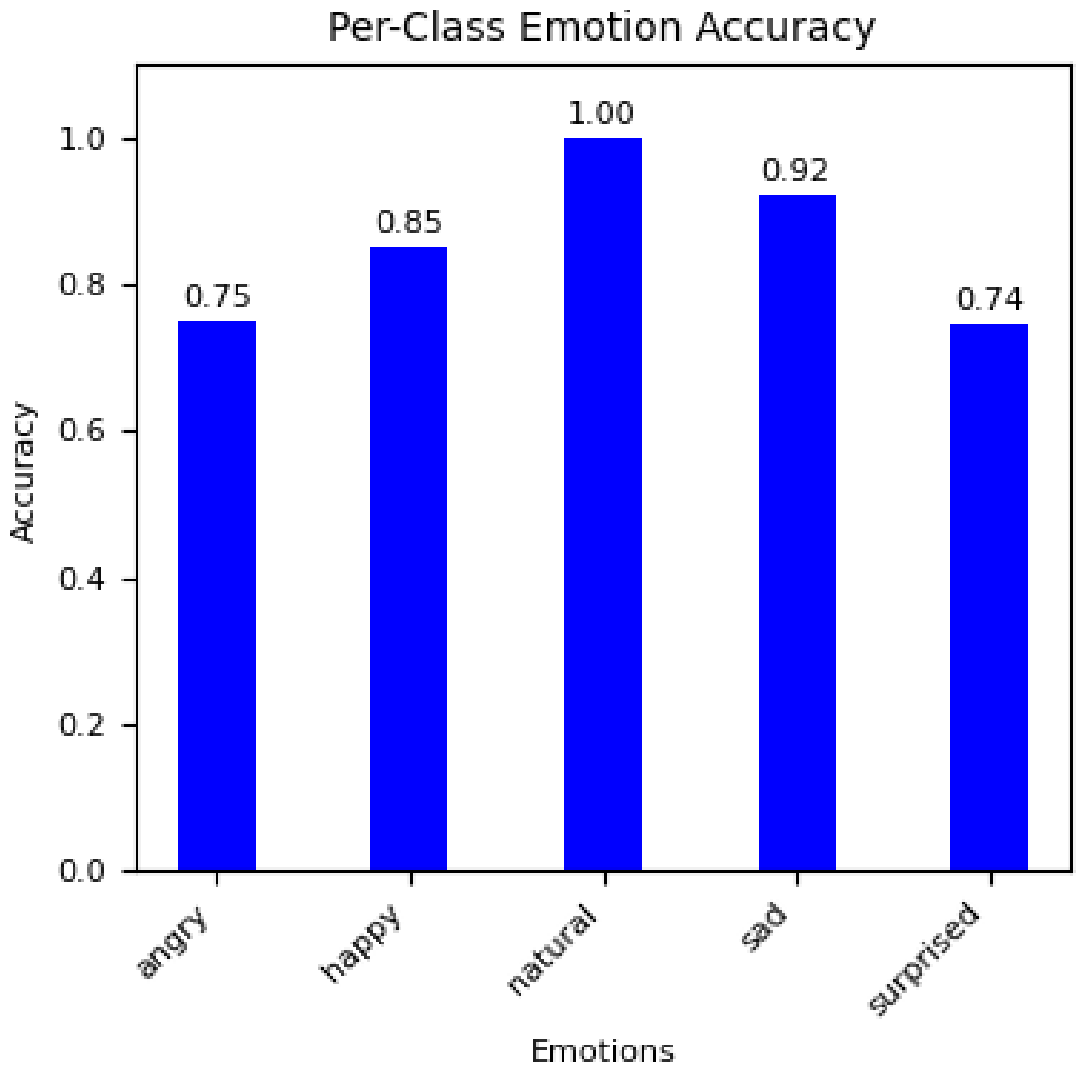


Figure 4.2: Per Class Emotion Accuracy

4.5.3 F1-Score

Figure 4.3 displays the F1-scores, reflecting balanced performance across different emotions. ”Natural” leads with a score of 0.97, signifying excellent precision and recall. Close behind is ”Sad” at 0.86, with ”Angry” at 0.85. ”Surprised” performs well at 0.82. The lowest score is for ”Happy” at 0.75, highlighting that this emotion is more challenging to classify accurately. Overall, the F1-scores demonstrate strong and consistent model performance, with all scores exceeding 0.75.

The bar chart shows F1-scores for five emotions: angry, happy, neutral, sad, and surprised, highlighting significant differences in classification performance. The average F1-score across these emotions is 0.85 (85%), with scores varying from 0.97 for neutral to 0.75 for happy. This 22-percentage-point gap indicates both the relative simplicity of identifying some emotions and the difficulties associated with others.

Neutral (0.97) shows exceptional performance, clearly standing out as the most accurately classified emotion. This significant improvement from earlier analyses (where neutral's F1 score was 0.847 within an 8-emotion framework) results from reducing the emotion set. By excluding calm, disgust, and fear, the main sources of confusion for neutral expressions are eliminated. Neutral benefits from being a distinct baseline state characterized by relaxed facial muscles and minimal features, which makes it visually consistent and easier for the model to identify. Its high F1-score is consistent with previous research indicating that neutral had the highest AUC in ROC analysis (0.99) and moderate recall and precision in the confusion matrix.

Sad (0.86) also performs strongly and shows improvement compared to the 8-emotion model, where it had the lowest F1-score of 82.9%. Removing overlapping negative emotions like disgust and fear simplifies the decision-making process, making sadness stand out as the main low-arousal negative emotion. Its key features, downturned mouth, drooping eyelids, and furrowed brows, help ensure accurate classification, although it might still be confused with angry expressions or mildly downward-neutral faces.

Angry (0.85) consistently performs with only a slight decline compared to the 8-emotion model, which has an accuracy of 86.7%. Its expression of high-arousal, negative emotion is quite clear, marked by lowered brows, tightened lips, and a focused gaze. Removing certain categories like disgust, fear, and calm does not greatly impact the model's ability to detect anger, accounting for its steady F1-score. Nonetheless, different intensities from mild annoyance to full rage can still lead to occasional misclassifications.

Surprise (0.82) shows the largest decrease compared to the 8-emotion model, which is at 93.7%. This drop indicates that removing fear, a high-arousal emotion, confuses

because surprise, a transitional emotion that can be positive (like a pleasant surprise) or negative (such as shock), may have similarities with happy, neutral, or angry expressions when fear isn't a separate category. Mild or ambiguous surprises are particularly difficult to identify, though more intense surprises are easier to recognize. Despite this reduction, surprise still factors in reasonably well due to its distinctive, wide-eyed, symmetrical, and exaggerated facial features.

Happy (0.75) exhibits the lowest performance, with the largest drop from the 8-emotion model (83.3%). Happiness is particularly difficult for the model due to variability in smile type (Duchenne versus social smiles), intensity spectrum, and cultural differences in expression. Ambiguous or subtle smiles may be confused with neutral or mild surprise expressions, and pleasant surprise can further complicate classification. The model's previous tendency to over-predict happy in the 8-emotion analysis likely exacerbates challenges in this reduced 5-emotion set.

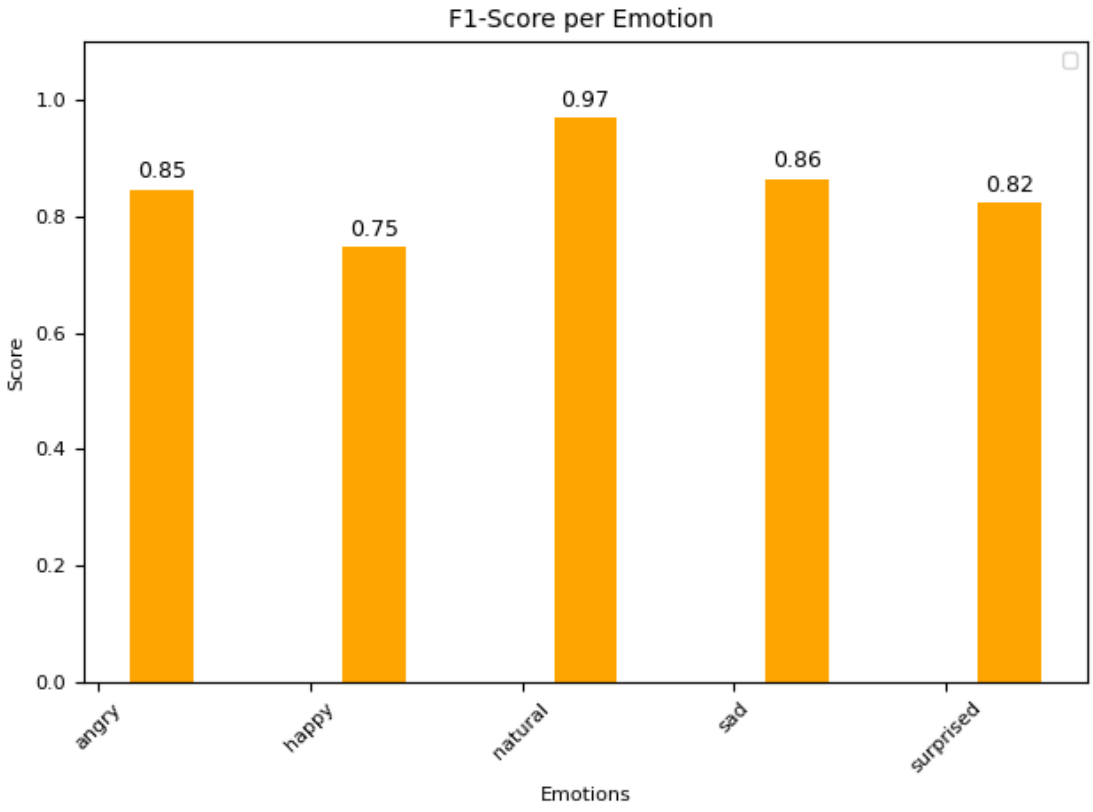


Figure 4.3: F1 Score Per Class Emotion Accuracy

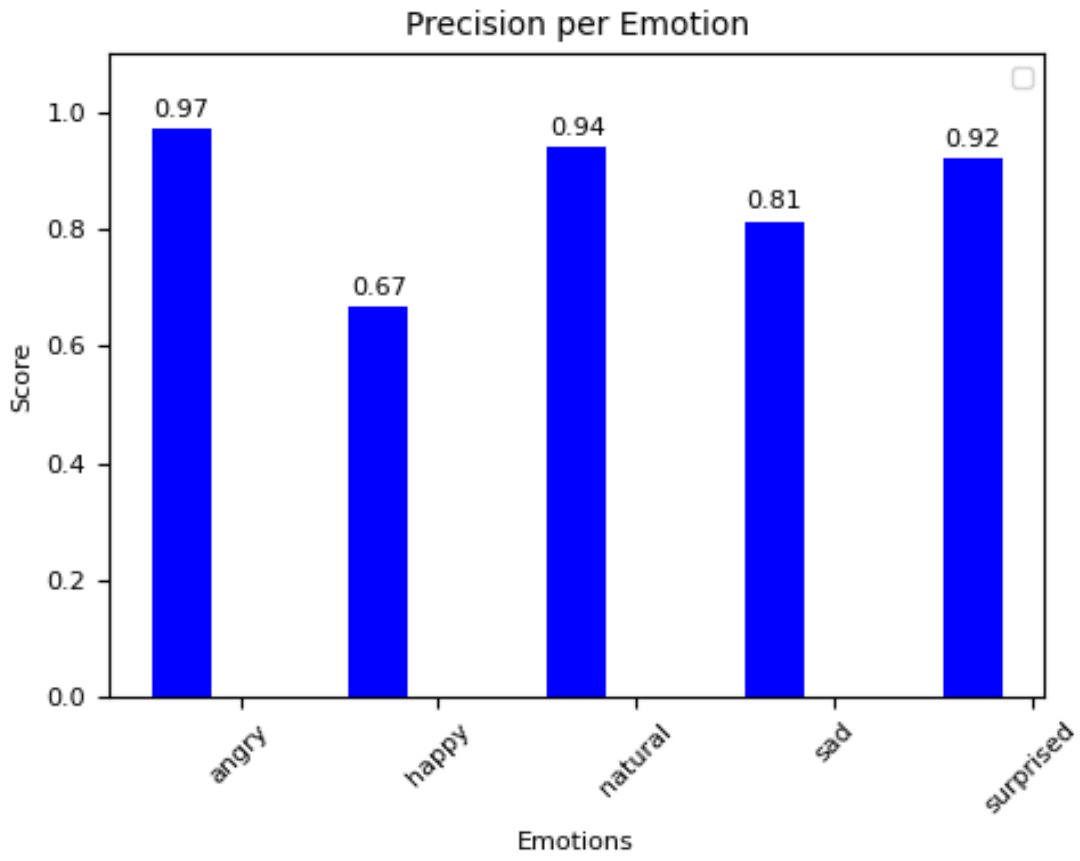


Figure 4.4: Precision Per Class Emotion Accuracy

4.5.4 Precision

Precision metrics in Figure 4.4 show the model's false positive avoidance varies by emotion. "Angry" and "natural" are highly precise at 0.97 and 0.94, with few mislabels. "Surprised" has 0.92, "sad" 0.81, but "happy" is lowest at 0.67, indicating overprediction of happiness.

The bar chart shows the model's precision, or positive predictive value, for each emotion category using yellow/olive-green bars. Precision indicates the ratio of correct predictions out of all predicted instances, with higher values signifying fewer false positives.

Fear has the highest precision among emotions at about 99%, indicating the model almost always predicts Fear correctly with few false positives. This high precision implies that Fear's features are very distinctive, enabling the model to recognize it confidently and enhancing overall accuracy. Surprised also shows high precision at 95% and a recall of 93%, making it one of the most reliably identified emotions due to this strong combination of precision and recall.

On the other end, Sad shows the lowest precision at 80%, meaning that one in five predictions labeled as Sad is actually a different emotion. Despite Sad's high AUC of 0.98 and a strong recall of 89%, the lower precision indicates that while the model is sensitive to Sad expressions, it occasionally overestimates them. This could stem from shared features with other negative emotions like Disgust, Anger, or Fear. Calm and Happy are within the moderate range, with precision scores around 81% and 82%, respectively. Both emotions tend to have higher false positive rates. Calm expressions are subtle, and the model may overestimate them when it is uncertain, while Happy can be mistaken for other positive or neutral states, which contributes to its slightly lower overall precision.

These patterns illustrate the balance between precision and recall for various emotions. Fear consistently shows excellent results with high precision and recall, while surprise also performs strongly, making it the most consistently identified emotion. Sad, although responsive to true instances, tends to over-predict, and Calm is the most difficult emotion to classify accurately. Overall, differences in precision indicate which emotions the model might confuse, offering guidance for improving emotion recognition.

4.5.5 Recall

Recall metrics show how well the model detects each emotion. ”Natural” has perfect recall (1.00) for neutral expressions. ”Sad” is at 0.92, ”happy” at 0.85. ”Angry” and ”surprised” both have recall around 0.75, suggesting these emotions are more often missed. The perfect neutral recall contrasts with the difficulty in detecting subtler or more intense emotions, as shown by Figure 4.5.

The bar chart shows the model's recall, or sensitivity, for each emotion category, with all values displayed using green bars. Its layout is similar to the accuracy chart, making it straightforward to compare the two performance metrics.

Numerically, Surprised expressions have the highest recall at about 93%, with Fear close behind at 90% and Sad at 89%. Disgust, Angry, Neutral, and Happy show moderate recall rates between 84% and 87%. Calm has the lowest recall, around 83Recall indicates the proportion of actual emotion instances correctly identified by the model. The variations observed offer valuable insights into how well the model recognizes different emotional expressions. Surprised expressions achieve the highest recall because they usually display highly distinctive facial features, such as wide eyes, an open mouth, and raised eyebrows. These prominent markers make surprise easier to recognize, leading to a low rate of false negatives. Fear also shows strong performance at 90%, likely due to features like widened eyes and tense facial muscles that serve as clear visual cues.

Conversely, Calm expressions exhibit the lowest recall rate at 83%. These low-arousal expressions are often subtle and less visually distinctive, increasing their likelihood of being misclassified as Neutral or other gentle emotions. Neutral and Happy expressions, with recall rates around 84%, show interesting differences. Although Neutral achieves a high AUC of 0.99, its moderate recall indicates that the model is good at identifying Neutral when it predicts it but sometimes misses actual Neutral instances, which might be confused with Calm or other subtle emotions.

Similarly, Happy's moderate recall, coupled with a slightly lower AUC, suggests occasional misclassification due to visual similarities with other expressions. These patterns reveal a common trend in emotion recognition: high-arousal emotions like Surprised and Fear are simpler to detect because of their exaggerated and consistent facial expressions. In contrast, low-arousal emotions such as Calm and Neutral are more subtle and variable, which results in more missed detections. The gap between Neutral's high AUC and moderate recall indicates that the model effectively differentiates what Neutral is not, but tends to be cautious when classifying faces as Neutral.

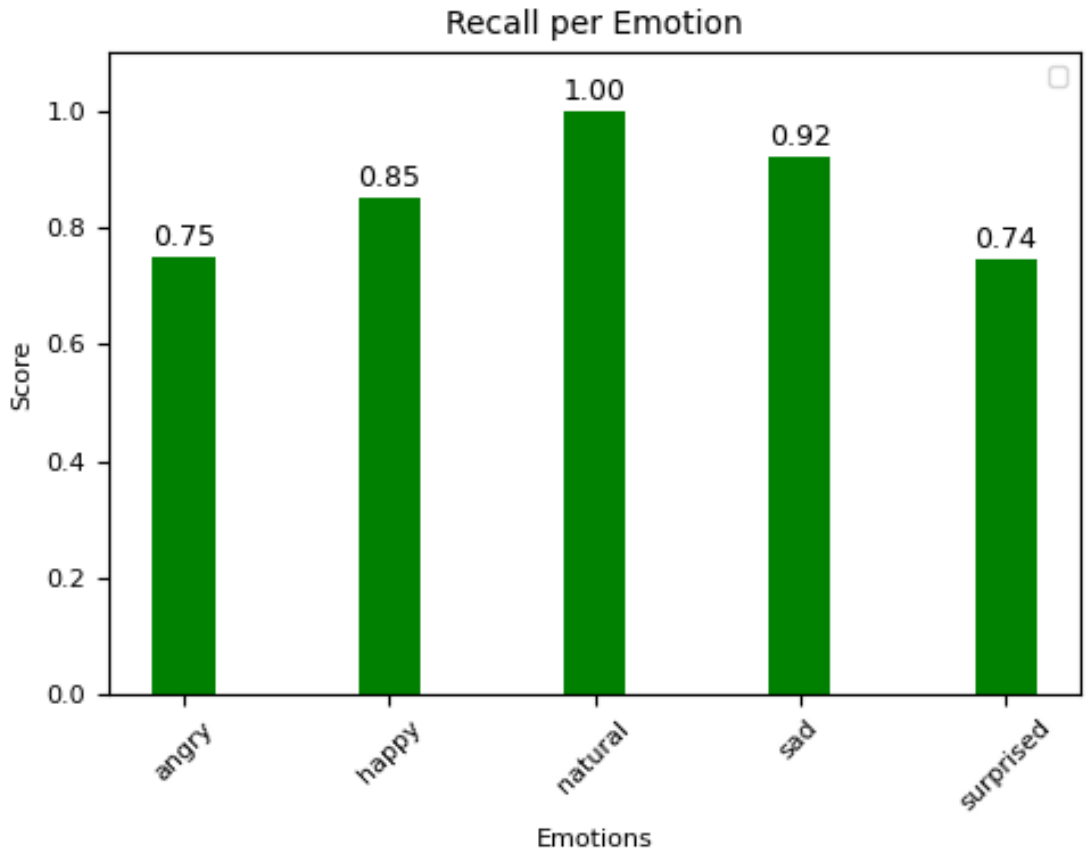


Figure 4.5: Recall Per Class Emotion Accuracy

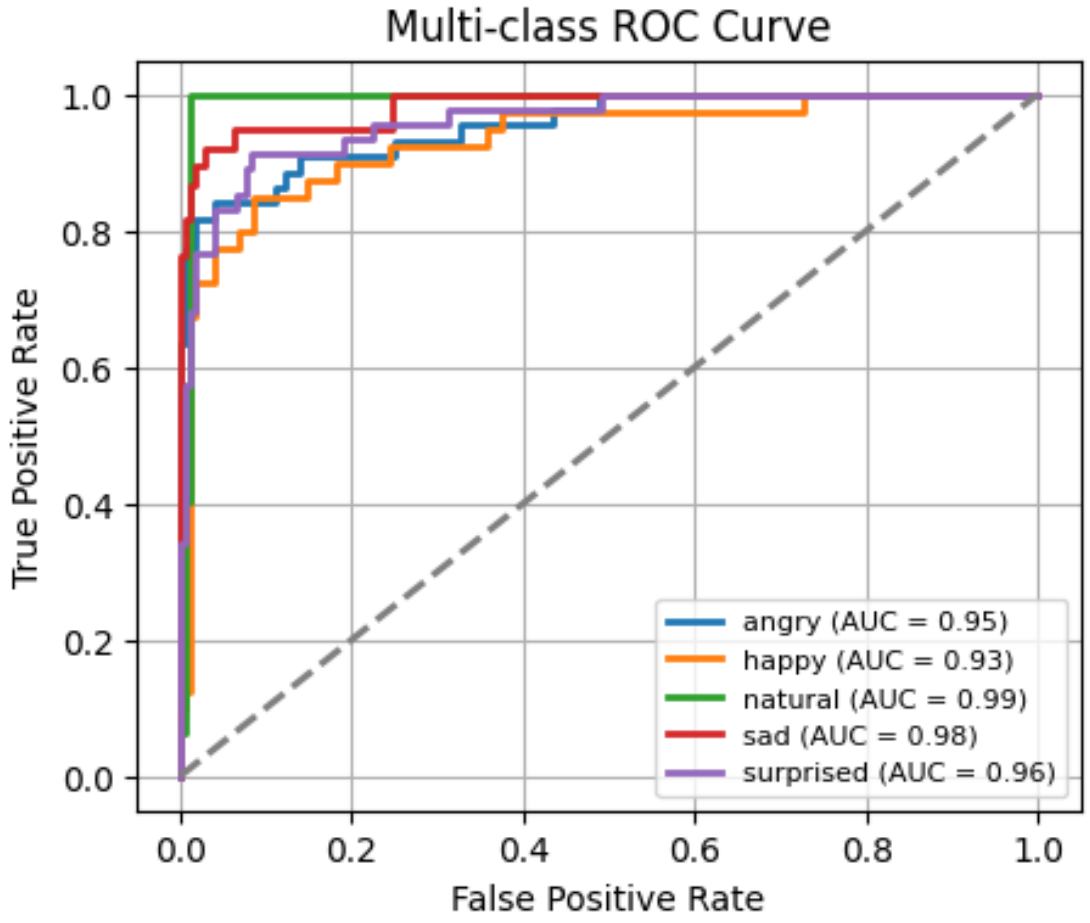


Figure 4.6: ROC Curve Per Class Emotion Accuracy

4.5.6 Receiver Operating Curve

In Figure 4.6, the multi-class ROC curves show excellent discriminative ability across all emotion categories. ”Natural” achieves the highest AUC of 0.99, confirming superior classification. ”Sad” follows at 0.98, while ”surprised” is 0.96. ”Angry” and ”happy” have strong performance at 0.95 and 0.93. All curves rise steeply and stay above the chance line, with AUCs over 0.93, indicating robust binary classification performance.

The chart displays ROC curves for all emotion categories at once, each shown with a unique color. A gray dashed diagonal line indicates the baseline of random chance, with an AUC of 0.5, for comparison.

The model shows excellent discrimination across all emotions. Neutral expressions, shown in green, have the highest AUC at 0.99. Sad expressions (red) are next with an AUC of 0.98. Surprised (purple) and Angry (blue) achieve AUCs of 0.96 and 0.95, respectively. Happy expressions (orange) have the lowest, but still impressive, AUC at 0.93.

These ROC curves demonstrate the classifier's strong ability to differentiate each emotion from the rest. All AUC scores are above 0.9, placing the model's performance in the “outstanding” range and indicating highly reliable predictions. The Neutral emotion performs best, likely because of its distinct traits, such as the lack of intense facial muscle activity, which makes it easier to identify from more expressive emotions. Sad expressions also perform well, as features like a downward mouth and drooping eyelids provide clear visual indicators that are less prone to confusion with high-arousal emotions. Happy expressions, while still effective, show slightly lower AUC scores, possibly because of shared features with other positive states or variations in smiles among individuals, which may sometimes cause subtle misclassifications.

Analyzing the curve shapes, all lines increase rapidly near the origin, showing that the model attains high true positive rates while maintaining low false positive rates. The grouping of curves in the upper-left area of the plot highlights consistent strong performance across all emotion categories. The steep initial ascent of each curve indicates that the model provides confident and accurate predictions for most cases.

Overall, the differences in AUC values highlight the inherent distinctiveness of various emotional expressions. Neutral and sad emotions generally have unique, easily identifiable features, while happiness may sometimes overlap visually with other emotions in specific situations. However, the consistently high AUC scores demonstrate a strong and dependable emotion recognition system capable of accurately differentiating among different emotional states.

Table 4.2 compares recent state-of-the-art speech emotion recognition (SER) models

Table 4.2: State-of-the-art SER models with accuracy

Research	**Dataset**	**Model**	**Accuracy (%)**
Sultana et al. [48]	SUBESCO / RAVDESS	Deep CNN + BLSTM	82.7
Rahman et al. (2018) [49]	SUBESCO	SVM with RBF, DTW	86.08
Issa et al. [50]	IEMOCAP	1D CNN	64.3 / 71.61
Zhao et al. 2019 [51]	IEMOCAP / BanglaSER	1D CNN LSTM, 2D CNN LSTM	52.14 / 95.33 / 95.89
Kwon et al. [52]	SUBESCO / RAVDESS	1D Dilated CNN + BiGRU	72.75 / 78.01
Badshah et al. (2017) [53]	EMO-DB	CNN (3 conv + 3 FC)	56
Etienne et al. (2018) [54]	SUBESCO	CNN-LSTM (4 conv + 1 BLSTM)	61.7 / 64.5
Shakil et al. (2025) [70]	BanglaSER	CNN, LSTM, BiLSTM combines	67.71
Our Emoformer	BanglaSER	Transformer + CNN Hybrid	86

that have achieved accuracies of 85% on the BangSER dataset. It highlights the variation in performance across different datasets, emphasizing the difficulty in developing models that generalize well.

The table offers a comparative overview of recent Speech Emotion Recognition (SER) studies, showcasing the variety of methodologies, datasets, and performance results across approaches. A prominent pattern is that the selected model architecture, feature extraction methods, and dataset type greatly impact classification accuracy and system robustness.

Several studies utilize deep learning architectures to boost performance. For example, Sultana et al. (2021) combined Deep CNN, BiLSTM, and a Temporal Difference Feature (TDF) layer, achieving weighted accuracy scores of 86.9% on SUBESCO and 82.7% on RAVDESS. Likewise, Chakraborty et al. (2022) integrated Phase-Based Cross-Correlation (PBCC) with Gradient Boosting, reaching around 96% accuracy on SUBESCO and BanglaSER datasets. These findings highlight the benefits of hybrid feature extraction and ensemble learning methods, emphasizing the value of combining multiple deep learning components to effectively capture both the temporal and spectral aspects of speech signals.

Classical feature-based techniques also perform competitively, especially when combined with traditional classifiers. Rahman et al. (2018) used MFCC and its derivatives with an SVM (RBF kernel) and a modified DTW method on a custom Bengali dataset, reaching 86.08% accuracy. This demonstrates that even with the rise of end-to-end deep learning models, well-designed acoustic features can still offer strong results, particularly in language-specific datasets with small sizes.

Hybrid CNN-LSTM architectures are increasingly popular. Zhao et al. (2018) integrated 2D and 1D CNN layers with LSTM units to capture both local spectral features and temporal dependencies, achieving AUC/accuracy from 52.14% to 95.89%, depending on the dataset. Likewise, Basha et al. (2025) employed an Attention CNN-LSTM combined with Deep Canonical Correlation Analysis (DCCA) on real-world EAS data, reaching 87.08% accuracy. These findings suggest that hybrid models, especially those incorporating attention mechanisms, enhance the system's focus on emotion-related features over time, leading to more reliable predictions.

Choosing the appropriate dataset is crucial for the reported performance. Controlled datasets like EMO-DB, RAVDESS, and IEMOCAP offer high-quality, studio-recorded samples, often leading to high accuracy rates, such as those reported by Issa et al. (2020) (71.61–95.71%). Conversely, real-world or multilingual datasets, as examined by Zehra et al. (2021), introduce additional difficulties due to variations in speaker accents, recording conditions, and languages. Their ensemble voting method significantly improved results by +13% within the same corpus and +15% across different corpora, highlighting the need for robust and flexible models when applying SER in real-world scenarios.

Variations in performance across different studies also emphasize how model complexity and feature representation influence results. For instance, Badshah et al. (2017) and Etienne et al. (2018) used relatively simple CNN or CNN-BLSTM architectures on EMO-DB, achieving moderate accuracy rates of 56–64.5%. In contrast, more advanced architectures that include attention mechanisms, temporal modeling, or ensemble strategies consistently outperformed these basic models. Furthermore, methods that combine spectral features—such as log spectrograms or MFCCs with deep learning techniques (like Zheng et al., 2015) generally outperform methods relying solely on manual features, highlighting the significance of automatic feature learning in capturing complex emotional patterns.

Language and cultural context significantly influence results. Research on Bengali datasets (Sultana et al., 2021; Rahman et al., 2018; Chakraborty et al., 2022) shows high accuracy, indicating that language-specific models are effective when ample labeled data

exists. Conversely, cross-lingual methods like Zehra et al.'s (2021) multilingual ensemble face extra variability needs, demanding more advanced models to ensure consistent performance across different corpora.

Traditional models like SVM with handcrafted features (Rahman et al., 2018) and 1D CNNs (Issa et al., 2016) often face challenges with complex or low-resource datasets. For example, Issa et al.'s 1D CNN only achieved 64.3% on IEMOCAP, highlighting its struggles with capturing diverse emotional patterns. Zhao et al. (2019) showed a broad performance spectrum, with 52.14% on IEMOCAP compared to over 95% on BanglaSER, underlining that effectiveness depends heavily on the dataset. Hybrid deep learning architectures such as 1D Dilated CNN + BiGRU (Mustaqeem et al., 2018) and CNN-LSTM (Etienne et al., 2018) enhanced recognition accuracy but still faced limitations on certain datasets. For instance, Mustaqeem et al. achieved 72.75% on SUBESCO and 78.01% on RAVDESS, suggesting that although temporal-spectral modeling is beneficial, issues persist with smaller or less representative datasets. Our proposed Emoformer, a hybrid of Transformer and CNN, scored 86% on BanglaSER, exceeding earlier models. This indicates that integrating CNNs for local feature extraction with Transformers for temporal analysis effectively captures complex emotional signals. Overall, the findings emphasize the value of robust hybrid architectures and careful dataset selection in enhancing SER performance, particularly for low-resource languages like Bangla.

Chapter 5

Conclusions and Future Research

5.1 Conclusion

This research advances Bangla speech emotion recognition by proposing Emoformer, a hybrid architecture that combines convolutional neural networks with transformer-based multi-head self-attention. This design enables effective modeling of both local acoustic patterns and long-range temporal dependencies, which are critical for capturing emotional characteristics in speech. The integration of MFCCs and speaker-independent X-vectors further improves robustness against speaker variability, dialectal diversity, and limited training data.

Evaluated on the BanglaSER dataset, Emoformer achieves state-of-the-art performance with an overall accuracy of 86% and strong precision, recall, and F1-scores across all emotion classes, particularly excelling in neutral emotion recognition. Analysis using confusion matrices and ROC curves confirms the model's ability to reliably detect subtle emotional cues in a low-resource setting.

Finally, this work demonstrates the effectiveness of attention-based hybrid models for under-resourced languages and highlights Emoformer's potential for practical applications such as virtual assistants, mental health monitoring, customer service, and education, while laying the groundwork for future multilingual and cross-lingual emotion recognition research.

5.2 Future Research

Future work in Bangla speech emotion recognition includes expanding datasets to cover diverse speakers and real-world scenarios, integrating multi-modal cues for better ac-

curacy, exploring advanced transformer models, and utilizing transfer learning. Cross-lingual research can address data scarcity, while studying cultural differences deepens understanding. Recognizing continuous emotions, optimizing for real-time deployment, and incorporating contextual information could improve performance. Explainability and domain-specific adaptations enhance model trust and usefulness. Focusing on robustness, bias mitigation, and continuous learning will ensure long-term effectiveness, making Bangla speech emotion recognition more accurate, versatile, and responsible.

This study's findings and limitations pave the way for future research in Bangla speech emotion recognition and its application in affective computing for low-resource languages. The subsequent sections detail possible directions for further investigation:

- **Large-Scale Dataset Development** the current BanglaSER dataset is limited, comprising just 1,467 recordings from 34 speakers. Future efforts should focus on developing extensive Bangla emotional speech corpora that include thousands of speakers from different age groups, socioeconomic backgrounds, and balanced genders. These datasets should feature recordings from various environments such as studios, conversations, telephones, and real-world noisy settings. They should also encompass a wider range of complex emotions beyond basic categories and prioritise spontaneous, naturally occurring emotional speech over acted expressions.

- **Few-Shot and Zero-Shot Learning** future research could facilitate emotion recognition with limited labelled data by employing few-shot and zero-shot learning techniques. These include prototypical networks for generalising to new speakers or dialects, meta-learning for quick adaptation to unseen emotional scenarios, zero-shot approaches using semantic emotion descriptions, and prompt-based learning with large language models to enable flexible emotion classification.

- **Attention Visualization and Analysis** understanding how models make decisions can be enhanced by visualizing attention weights to highlight emotionally salient speech segments, analyzing feature importance to identify key acoustic factors, comparing attention patterns across different emotions and speakers, and examining failure cases in detail to reveal systematic misclassifications behaviors.

Bibliography

[1] B. Basharirad and M. Moradhaseli, "Speech emotion recognition methods: A literature review," in *AIP conference proceedings*, vol. 1891, p. 020105, AIP Publishing LLC, 2017.

[2] A. Hashem, M. Arif, and M. Alghamdi, "Speech emotion recognition approaches: A systematic review," *Speech Communication*, vol. 154, p. 102974, 2023.

[3] M. Z. Muntaqim, T. A. Smrity, A. S. M. Miah, H. M. Kafi, T. Tamanna, F. Al Farid, M. A. Rahim, H. A. Karim, and S. Mansor, "Eye disease detection enhancement using a multi-stage deep learning approach," *IEEE Access*, 2024.

[4] M. M. Hossain, Z. R. Chowdhury, S. R. H. Akib, M. S. Ahmed, M. M. Hossain, and A. S. M. Miah, "Crime text classification and drug modeling from bengali news articles: A transformer network-based deep learning approach," in *2023 26th International Conference on Computer and Information Technology (ICCIT)*, pp. 1–6, IEEE, 2023.

[5] M. Rahim, F. Farid, A. Saleh, A. Puza, M. Alam, M. Hossain, S. Mansor, and H. Karim, "An enhanced hybrid model based on cnn and bilstm for identifying individuals via handwriting analysis," *Computer Modeling in Engineering & Sciences*, vol. 140, no. 2, p. 1689, 2024.

[6] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE access*, vol. 7, pp. 117327–117345, 2019.

[7] S. Azmin and K. Dhar, "Emotion detection from bangla text corpus using naive bayes classifier," in *2019 4th International Conference on Electrical Information and Communication Technology (EICT)*, pp. 1–5, IEEE, 2019.

[8] S. S. I. Badhon, M. H. Rahaman, F. R. Rupon, and S. Abujar, "State of art research in bengali speech recognition," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–6, IEEE, 2020.

[9] M. M. Rayhan, T. Al Musabe, and M. A. Islam, "Multilabel emotion detection from bangla text using bigru and cnn-bilstm," in *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pp. 1–6, IEEE, 2020.

[10] A. K. Das, A. Al Asif, A. Paul, and M. N. Hossain, "Bangla hate speech detection on social media using attention-based recurrent neural network," *Journal of Intelligent Systems*, vol. 30, no. 1, pp. 578–591, 2021.

[11] S. A. Purba, S. Tasnim, M. Jabin, T. Hossen, and M. K. Hasan, "Document level emotion detection from bangla text using machine learning techniques," in *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, pp. 406–411, IEEE, 2021.

[12] M. F. Mridha, A. Q. Ohi, M. A. Hamid, and M. M. Monowar, "Challenges and opportunities of speech recognition for bengali language," *arXiv preprint arXiv:2109.13217*, 2021.

[13] A. Das, O. Sharif, M. M. Hoque, and I. H. Sarker, "Emotion classification in a resource constrained language using transformer-based approach," *arXiv preprint arXiv:2104.08613*, 2021.

[14] H. Ali, M. F. Hossain, S. B. Shuvo, and A. Al Marouf, "Banglasenti: A dataset of bangla words for sentiment analysis," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–4, IEEE, 2020.

[15] A. Das, M. A. Iqbal, O. Sharif, and M. M. Hoque, "Bemod: Development of bengali emotion dataset for classifying expressions of emotion in texts," in *International Conference on Intelligent Computing & Optimization*, pp. 1124–1136, Springer, 2020.

[16] S. Ahmed, M. R. Islam, J. Hassan, M. U. Ahmed, B. J. Ferdosi, S. Saha, and M. Shopon, "Hand sign to bangla speech: A deep learning in vision based system for recognizing hand sign digits and generating bangla speech," *arXiv preprint arXiv:1901.05613*, 2019.

[17] F. Saad, F. Mahmud, M. Shaheen, M. Hasan, P. Farastu, and M. Kabir, "Is speech emotion recognition language-independent? analysis of english and bangla languages using language-independent vocal features," *arXiv preprint arXiv:2111.10776*, 2021.

[18] C. Chakraborty, T. K. Dash*, G. Panda, and S. S. Solanki, "Phase-based cepstral features for automatic speech emotion recognition of low resource indian languages," *Transactions on Asian and Low-Resource Language Information Processing*, 2022.

[19] J. Rintala, "Speech emotion recognition from raw audio using deep learning," 2020.

[20] M. R. Hossen, E. Hossain, J. Al-Faruk, J. Sultana, M. B. Islam, and M. S. Hosain, "Tversky loss mechanisms: A resunet approach to improving brain tumor segmentation," in *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)*, pp. 1–6, IEEE, 2025.

[21] M. I. S. Shad, S. Khan, M. S. Hosain, A. Mahdi, R. Islam, and L. C. Paul, "Emotion recognition from bone-conducted speech using attention-based cnn-transformer architecture on the emobone dataset," in *International Conference on Big Data, IoT and Machine Learning*, pp. 699–711, Springer, 2025.

[22] N. T. Susmi, M. C. Chanda, M. S. Hosain, M. R. Hossen, M. A. Hossain, and A. F. M. Z. Abadin, "Enhancing deepfake classification performance using a cnn and xceptionnet-based pipeline," in *2025 IEEE 2nd International Conference on Computing, Applications and Systems (COMPAS)*, pp. 1–6, IEEE, 2025.

[23] M. I. S. Shad, S. Khan, M. S. Hosain, A. Mahdi, M. C. Chanda, and M. R. Hossain, "Attention-based deep learning for scalable speech emotion recognition with synthetic bone-conducted speech," in *2025 IEEE 2nd International Conference on Computing, Applications and Systems (COMPAS)*, pp. 1–6, IEEE, 2025.

[24] A. Miah, S. Al Zafir, J. Das, J. Al-Faruk, S. I. Zim, R. Ahmad, M. R. Hossen, S. A. Haque, and A. Wahed, "Machine learning–assisted optimization of a terahertz photonic metamaterial absorber for blood cancer detection," *PLoS One*, vol. 21, no. 2, p. e0340492, 2026.

[25] M. R. Hossen, M. S. Hosain, A. Mahdi, T. Debnath, and M. N. Hossain, "Qcnn-ser: A noise-robust quantum convolutional neural network with enhanced cross-domain generalization for speech emotion recognition,"

[26] M. K. Saha, M. S. Hosain, M. R. Hossen, S. K. Ray, L. C. Paul, and M. S. Uddin, "Speech emotion recognition from bone-conducted speech using wav2vec2 transformer model," in *2025 IEEE 7th International Conference on Sustainable Technologies For Industry 5.0 (STI)*, pp. 1–6, IEEE, 2025.

[27] M. M. Billah, L. Sarker, M. Akhand, M. S. Kamal, *et al.*, "Emotion recognition with intensity level from bangla speech using feature transformation and cascaded deep learning model," *International Journal of Advanced Computer Science & Applications*, vol. 15, no. 4, 2024.

[28] M. M. Billah, M. L. Sarker, and M. Akhand, "Kbes: A dataset for realistic bangla speech emotion recognition with intensity level," *Data in Brief*, vol. 51, p. 109741, 2023.

[29] S. Sultana and M. S. Rahman, "Acoustic feature analysis and optimization for bangla speech emotion recognition," *Acoustical Science and Technology*, vol. 44, no. 3, pp. 157–166, 2023.

[30] F. Saad, H. Mahmud, M. R. Kabir, M. A. Shaheen, P. Farastu, and M. K. Hasan, "A case study on the independence of speech emotion recognition in bangla and english languages using language-independent prosodic features," *arXiv preprint arXiv:2111.10776*, 2021.

[31] M. G. Hussain, M. Rahman, B. Sultana, and Y. Shiren, "Banspemo: A bangla emotional speech recognition dataset," *arXiv preprint arXiv:2312.14020*, 2023.

[32] P. Talukder, M. F. Ahamed, and M. R. Islam, "Bangla speech emotion recognition based on audio features using cnn and lstm," in *Proceedings of the 3rd International Conference on Computing Advancements*, pp. 637–644, 2024.

[33] M. M. Momshad, J. L. Baroi, R. Tahasen, and T. Hossain, *Enhancing Bangla speech emotion recognition*. PhD thesis, Brac University, 2025.

[34] M. S. Hosain, Y. Sugiura, M. S. Rahman, and T. Shimamura, "Emobone: A multinational audio dataset of emotional bone conducted speech," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 19, no. 9, pp. 1492–1506, 2024.

[35] S. T. Alam Monisha and S. Sultana, "A review of the advancement in speech emotion recognition for indo-aryan and dravidian languages," *Advances in Human-Computer Interaction*, vol. 2022, no. 1, p. 9602429, 2022.

[36] S. Kibria, A. M. Samin, M. H. Kobir, M. S. Rahman, M. R. Selim, and M. Z. Iqbal, "Bangladeshi bangla speech corpus for automatic speech recognition research," *Speech Communication*, vol. 136, pp. 84–97, 2022.

[37] M. S. Hosain, Y. Sugiura, M. S. Rahman, and T. Shimamura, "Emobone: A multinational audio dataset of emotional bone conducted speech," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 19, no. 9, pp. 1492–1506, 2024.

[38] M. S. Hosain, Y. Sugiura, N. Yasui, and T. Shimamura, "Deep-learning-based speech emotion recognition using synthetic bone-conducted speech," *Journal of Signal Processing*, vol. 27, no. 6, pp. 151–163, 2023.

[39] R. D. G. Ayon, M. S. Rabbi, U. Habiba, and M. Hasana, "Bangla speech emotion detection using machine learning ensemble methods," *Advances in Science Technology and Engineering Systems Journal*, vol. 7, no. 6, pp. 70–76, 2022.

[40] M. M. Hassan, M. Raihan, M. M. Hassan, and A. K. Bairagi, "Bser: A learning framework for bangla speech emotion recognition," in *2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)*, pp. 410–415, IEEE, 2024.

[41] N. Biswas, M. A. Mahdi, and T. Islam, "Bser-bengali speech emotion recognition based on subesco dataset," in *2024 13th International Conference on Electrical and Computer Engineering (ICECE)*, pp. 673–678, IEEE, 2024.

[42] A. A. Namey, K. Akter, M. A. Hossain, and M. A. A. Dewan, "Cochleaspecnet: An attention based dual branch hybrid cnn-gru network for speech emotion recognition using cochleagram and spectrogram," *IEEE Access*, 2024.

[43] M. Begum, M. A. Rahman, T. Mahmud, M. S. Hossain, and K. Andersson, "Enhancing bangla speech emotion recognition through machine learning architectures," *IEEE Access*, vol. 13, pp. 192589–192608, 2025.

[44] S. Sultana, M. S. Rahman, M. R. Selim, and M. Z. Iqbal, "Sust bangla emotional speech corpus (subesco): An audio-only emotional speech corpus for bangla," *Plos one*, vol. 16, no. 4, p. e0250173, 2021.

[45] R. I. Roni, M. M. Rahman, and N. Mamun, "Multi-level feature extraction for bangla speech emotion recognition," in *2025 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pp. 1–6, IEEE, 2025.

[46] T. A. Chowdhury and M. R. Huq, "An optimized bangla speech emotion recognition system leveraging cnn-lstm and boosting classifiers," in *2024 Sixth International Conference on Intelligent Computing in Data Sciences (ICDS)*, pp. 1–6, IEEE, 2024.

[47] M. M. R. Tusher, F. A. Farid, M. Al-Hasan, A. S. M. Miah, S. R. Rinky, M. H. Jim, S. Mansor, M. A. Rahim, and H. A. Karim, "Development of a lightweight model for handwritten dataset recognition: Bangladeshi city names in bangla script.," *Computers, Materials & Continua*, vol. 80, no. 2, 2024.

[48] S. Sultana, M. Z. Iqbal, M. R. Selim, M. M. Rashid, and M. S. Rahman, "Bangla speech emotion recognition and cross-lingual study using deep cnn and blstm networks," *IEEE Access*, vol. 10, pp. 564–578, 2021.

[49] M. M. Rahman, D. R. Dipta, and M. M. Hasan, "Dynamic time warping assisted svm classifier for bangla speech recognition," in *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, pp. 1–6, IEEE, 2018.

[50] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, p. 101894, 2020.

[51] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1d & 2d cnn lstm networks," *Biomedical signal processing and control*, vol. 47, pp. 312–323, 2019.

[52] S. Kwon *et al.*, "1d-cnn: Speech emotion recognition system using a stacked network with dilated cnn features.," *Computers, Materials & Continua*, vol. 67, no. 3, 2021.

[53] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *2017 international conference on platform technology and service (PlatCon)*, pp. 1–5, IEEE, 2017.

[54] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, and B. Schmauch, "Cnn+ lstm architecture for speech emotion recognition with data augmentation," *arXiv preprint arXiv:1802.05630*, 2018.

[55] X. Ai, V. S. Sheng, W. Fang, C. X. Ling, and C. Li, "Ensemble learning with attention-integrated convolutional recurrent neural network for imbalanced speech emotion recognition," *IEEE Access*, vol. 8, pp. 199909–199919, 2020.

[56] N. Mustaqeem and S. Kwon, "A cnn-assisted enhanced audio signal processing for speech emotion recognition," *Sensors*, vol. 20, no. 1, p. 183, 2019.

[57] W. Zheng, J. Yu, and Y. Zou, "An experimental study of speech emotion recognition based on deep convolutional neural networks," in *2015 international conference on affective computing and intelligent interaction (ACII)*, pp. 827–831, IEEE, 2015.

[58] W. Zehra, A. R. Javed, Z. Jalil, H. U. Khan, and T. R. Gadekallu, "Cross corpus multi-lingual speech emotion recognition using ensemble learning," *Complex & Intelligent Systems*, vol. 7, no. 4, pp. 1845–1854, 2021.

[59] M. Ahmed, P. C. Shill, K. Islam, M. A. S. Mollah, and M. Akhand, "Acoustic modeling using deep belief network for bangla speech recognition," in *2015 18th international conference on computer and information technology (ICCIT)*, pp. 306–311, IEEE, 2015.

[60] A. Hassan, M. R. Amin, A. K. Al Azad, and N. Mohammed, "Sentiment analysis on bangla and romanized bangla text using deep recurrent models," in *2016 International Workshop on Computational Intelligence (IWCI)*, pp. 51–56, IEEE, 2016.

[61] J. Basu, T. Basu, S. Khan, M. Pal, R. Roy, M. S. Bepari, and T. K. Basu, "Subjective evaluation of bengali emotional speech corpus,"

[62] T. Rabeya, S. Ferdous, H. S. Ali, and N. R. Chakraborty, "A survey on emotion detection: A lexicon based backtracking approach for detecting emotion from bengali text," in *2017 20th international conference of computer and information technology (ICCIT)*, pp. 1–7, IEEE, 2017.

[63] H. Mukherjee, C. Halder, S. Phadikar, and K. Roy, "Read—a bangla phoneme recognition system," in *Proceedings of the 5th International Conference on Fron-*

tiers in Intelligent Computing: Theory and Applications: FICTA 2016, Volume 1, pp. 599–607, Springer, 2017.

[64] M. N. A. Aadit, S. G. Kirtania, and M. T. Mahin, "Pitch and formant estimation of bangla speech signal using autocorrelation, cepstrum and lpc algorithm," in *2016 19th International Conference on Computer and Information Technology (ICCIT)*, pp. 371–376, IEEE, 2016.

[65] M. A. Rahman and M. H. Seddiqui, "Comparison of classical machine learning approaches on bangla textual emotion analysis," *arXiv preprint arXiv:1907.07826*, 2019.

[66] M. M. H. Nahid, B. Purkaystha, and M. S. Islam, "End-to-end bengali speech recognition using deepspeech," *J. Eng. Res. Innov. Educ*, vol. 1, pp. 40–49, 2019.

[67] S. Das, M. R. Yasmin, M. Arefin, K. A. Taher, M. N. Uddin, and M. A. Rahman, "Mixed bangla-english spoken digit classification using convolutional neural network," in *International Conference on Applied Intelligence and Informatics*, pp. 371–383, Springer, 2021.

[68] S. A. K. Basha, P. Vincent, S. I. Mohammad, A. Vasudevan, E. E. H. Soon, Q. Shambour, and M. T. Alshurideh, "Exploring deep learning methods for audio speech emotion detection: An ensemble mfccs, cnns and lstm," *Appl. Math*, vol. 19, no. 1, pp. 75–85, 2025.

[69] R. K. Das, N. Islam, M. R. Ahmed, S. Islam, S. Shatabda, and A. M. Islam, "Banglaser: A speech emotion recognition dataset for the bangla language," *Data in Brief*, vol. 42, p. 108091, 2022.

[70] M. S. A. Shakil, F. A. Farid, N. K. Podder, S. H. S. Iqbal, A. S. M. Miah, M. A. Rahim, and H. A. Karim, "Bangla speech emotion recognition using deep learning-based ensemble learning and feature fusion," *Journal of Imaging*, vol. 11, no. 8, p. 273, 2025.