TriVashi: A Multi-Stage System for Dialectal Speech Translation, Integrating Identification, Transcription, and Synthesis for Under-Resourced

By
Salauddin Majumder
212-15-4227

Meheron Nesa Surovi
212-15-4215

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements forthe **Degree of Bachelor of Science in Computer Science and Engineering**

Supervised by
Md. Sadekur Rahman
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by
Sharun Akter Khushbu
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
Dhaka, Bangladesh

September 16, 2025

APPROVAL

This Project titled "**TriVashi: A Multi-Stage System for Dialectal Speech Translation, Integrating Identification, Transcription, and Synthesis for Under-Resourced**," submitted by **Salauddin Majumder, ID: 212-15-4227** and **Meheron Nesa Surovi, ID: 212-15-4215** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **16-09-2025**.

BOARD OF EXAMINERS

Dr. S.M Aminul Haque — **Chairman**
Professor and Associate Head
Department of Computer Science and Engineering, FSIT,
Daffodil International University

Dr. Md Alamgir Kabir — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering, FSIT,
Daffodil International University

Mr. Md Assaduzzaman — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering, FSIT,
Daffodil International University

Dr. Md. Zulfiker Mahmud — **External Examiner**
Professor
Department of Computer Science and Engineering,
Jagannath University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Md. Sadekur Rahman**, **Assistant Professor**, Department of Computer Science and Engineering, Daffodil International University. We also declare that neither this projectnor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

15.9.25

Md. Sadekur Rahman
Assistant Professor
Department of Computer Science and Engineering Daffodil International University

Co-Supervised by:

15.9.2025

Sharun Akter Khushbu
Assistant Professor
Department of Computer Science and Engineering Daffodil International University

Submitted by:

16.09.25

Salauddin Majumder
Student ID:212-15-4227
Department of Computer Science and Engineering
Daffodil International University

16.09.25

Meheron Nesa Surovi
Student ID:212-15-4215
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGEMENTS

This work would not have been possible without the support and contributions of many individuals over the past two semesters. We are deeply grateful to everyone who has assisted us in one way or another.

First, we express our heartfelt thanks and gratefulness to the almighty for His divine blessing making it possible for us to complete the **Final Year Design Project(FYDP)** successfully.

We are grateful and wish our profound indebtedness to **Md. Sadekur Rahman** , **Assistant Professor**, Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Deep knowledge and keen interest of our supervisor in the field of **Natural Language Processing** to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting themat all stages have made it possible to complete this project.

We would like to express our heartfelt gratitude to the Head of the Department of Computer Science and Engineering, for his kind help in finishing our project and also to other faculty members and the staff of the Department of Computer Science and Engineering, Daffodil International University.

We would like to thank our entire course-mates at Daffodil International University, who took part in this discussion while completing the coursework.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

This increasing sophistication in natural language processing has in a paradoxical manner increased the digital divide and left thousands of low-resource languages and dialects underserved by technology. TriVashi is the first end-to-end, multi-stage speech-to-speech translator of the under-resourced dialect of the Noakhali, Sylheti, and Chittagong dialect of Bangladesh, where no integrated solution previously existed, and this research directly challenges this issue of digital linguistic inequality. The main outcome of this work is the establishment and the publication of a new, gender-balanced, 15,006-sample audio and parallel text corpus as an exemplary basis that is likely to stimulate future innovation. It uses a four-stage cascaded architecture, based on a proposed system identifies the dialect through a new visual-analytic system; this method re-frames the problem as an image classification challenge by transforming audio to Mel spectrograms and using a pre-trained DenseNet121-SVM classifier with the best 92.7%accuracy. After detection, the audio is sent to dialect specific Automatic Speech Recognition (ASR) and Neural Machine Translation (NMT) models. The experimental findings confirm the transformative effectiveness of transfer learning; with fine-tuning of large pre-trained models (Whisper-Small and BanglaT5) on the curated dialectal data, the ASR Word Error Rate (WER) dropped monumentally, starting at more than 289.2 percent, down to as low as 3.0, and NMT BLEU scores surged to 56.3. The new state-of-the-art standards set in this work are accompanied by a strong and replicable methodological template that confirms the small data, big model paradigm as a viable way of facilitating digital equity to under-resourced languages around the world.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter introduces the fundamental problem domain of digital linguistic inequality, with a specific focus on the under-resourced dialects of Bangladesh. It articulates the scientific and societal motivations driving this research, delineates the specific objectives of the study, and provides a high-level overview of the proposed methodological framework. Furthermore, it outlines the anticipated scholarly and practical outcomes and concludes by presenting the organizational structure of the report.

1.1 Introduction

The modern digital age is defined by the high level of paradox. Although the progress of natural language processing (NLP) has transformed the human-computer interaction, the gains of this technological tidal wave have been skewed to a few high-resource languages. It has resulted in a major digital language gap with thousands of languages and dialects which are used by billions of people being left with little or no technological service and at risk of being pushed out of digital use. Such disparity creates obstacles of immense proportions to information access, economic opportunity, and cultural expression to large portions of the world population [1].

This study lies in this world problem by setting the centre on the plentiful and multifaceted linguistic environment of Bangladesh. Although Standard Bangali is the official language in the country, a large percentage of its population is able to speak their own regional languages. This paper addresses in particular three of the most heightened but technologically underprivileged dialects, namely Chittagong, Sylhet, and Noakhali. The linguistic difference between the two varieties and Standard Bangla poses a serious communication problem that is made worse by the fact that there is virtually no strong technological solution to these communication problems [2].

The main issue discussed here is that there is no end-to-end framework, with any overall, integrated systematized benchmarking, to translate the speech of these regional dialects into Standard Bangla speech. It can be seen in the review of the existing literature that whereas advances were achieved concerning isolated components, including (Noakhali-specific dialect) automatic speech recognition (ASR), a significant research lacuna remains. No system has been developed that integrates the various stages required in identifying the dialect, and specialized speech-to-text conversion followed by the text-to-text translation in this particular multi-dialect scenario. Such technological gap is a major deterrent to scientific progress in Bangla NLP, as well as to the creation of useful, socially valuable applications.

The present condition of language technology can be viewed as an exemplary critical last mile issue with the implementation of artificial intelligence. Existing paradigms such as Whisper show outstanding abilities in standard as well as high resource

languages but collapse disastrously when faced with phonetic and lexical variations of regional dialects, with initial error rates of over 289.2% being recorded. The problem, therefore, is not only the issue of technology invention but of advanced adaptation and assimilation. This study thus transcends the evolution of generalist technology to the scaled-up and more meaningful scientific inquiry of how best to demystify and apply these mighty instruments to the particular linguistic requirements of the disadvantaged populations, and thus realize the potential of technology as something truly available and useful [3].

1.2 Motivation

Scientific Motivation

The scientific motivations are centered on addressing core challenges in the field of computational linguistics for under-resourced languages.

Addressing the Data Scarcity Bottleneck: The most formidable obstacle in low-resource NLP is the chronic scarcity of high-quality, large-scale, and well-annotated datasets [4]. Existing corpora for the target dialects are acknowledged to be severely deficient. A primary scientific motivation of this work is therefore to overcome this bottleneck through systematic data curation. The development of a novel, gender-balanced corpus comprising 15,006 audio samples with parallel texts is not treated as a mere prerequisite but as a core scientific contribution designed to be a foundational resource that will catalyze future research in Bangla language technology.

Innovating in Dialect Identification: This study is driven by the fact that the possibility of innovation in a new, cross-domain dialect identification methodology exists. The hypothesis that the acoustic properties of the speech can be successfully simulated and categorized in the visual domain lies in the basis of the study. This entails converting one-dimensional audio signals into two-dimensional Mel spectrograms and using the powerful feature extraction properties of trained Convolutional Neural Networks (CNNs), e.g. DenseNet121 and InceptionV3. It was motivated by the urge to go beyond classic acoustic feature engineering (i.e., MFCCs) and to test the transferability of rich visual features, initially trained on natural images, into the spectro-temporal code of human speech.

Exploring the Frontiers of Transfer Learning: This study is primarily driven by a wish to quantify strictly the effectiveness of the small data, big model paradigm in a severe low-resource setting. It attempts to answer a critical scientific question: To what degree can huge, multilingual models such as Whisper-Small and language specific models such as BanglaT5 be successfully scaled to highly specific and non-standard linguistic varieties? The rationale is to objectively quantify the performance difference between the pre-trained and fine-tuned condition of these models, thus providing strong empirical evidence of the strength and the efficacy of transfer learning as a dominant approach to establishing the technology of the linguistic periphery of the world.

Societal Motivation

The social drivers are based on the values of digital inclusion, cultural continuity, and facilitation of high-impact, real-life use.

Promoting Digital Inclusion and Accessibility: The overall rationale in society is the

idea of ensuring that millions of native speakers of Chittagong, Sylheti, and Noakhali experience digital equity. This technology is developed to break the communication barrier, thus allowing users of these dialects to enjoy digital services, education opportunities and economic opportunities in their vernacular language.

Conserving Linguistic Heritage: This study will play a direct role in the preservation and renewal of the Bangladesh rich linguistic heritage by developing digital applications that identify, process, and certify the regional dialects. It offers the technological basis on which applications that honor and maintain regional linguistic identities can be produced instead of encouraging their assimilation or disappearance in the digital world.

Enabling High-Impact Applications: The driving force is the possibility of transformative societal impact. An efficient and precise speech-to-speech translation system can be used as the foundation of the localized service in the sectors of high priority. This covers healthcare, where it can provide helpful communication between patients and medical workers; education, where it can enable development of learning materials in local languages; and media, where it can drive automated subtitling, settings of content localisation and accessibility features.

1.3 Objectives

In order to fill the research gaps identified, as well as meet the motivations stated, the current study has a series of specific, measurable, and ambitious objectives. The following are the objectives:

Corpus Curation: To design, develop, and deploy a mobile application for the systematic collection of a novel, large-scale, gender-balanced audio and parallel text corpus for the Noakhali, Sylheti, and Chittagong dialects, and to prepare this corpus for public release to the global research community.

Multi Feature Extractors & Systematic Benchmarking of Classifiers: To design and implement a novel, hybrid dialect identification system that conceptualizes the task as an image classification problem by converting audio signals to Mel spectrograms and leveraging pre-trained CNNs (DenseNet121, InceptionV3, ResNet101, VGG19) as feature extractors. Also conduct a rigorous, comparative analysis of multiple machine learning classifiers—including Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Naive Bayes—to identify the optimal model for classifying the high-dimensional features extracted by the CNN architectures.

Development of High-Accuracy Dialectal ASR: To develop and fine-tune three distinct, high-accuracy Automatic Speech Recognition (ASR) models, one for each target dialect, by adapting the pre-trained Whisper-Small architecture on the curated dialect-specific datasets.

Development of Dialect-to-Standard Text Translation: To develop and fine-tune three effective text-to-text translation models capable of converting the transcribed regional text into Standard Bangla, utilizing the language-specific BanglaT5 transformer architecture.

End-to-End Framework Integration and Evaluation: To integrate the optimized

components into a cohesive, multi-stage speech-to-speech translation pipeline and to establish new state-of-the-art performance benchmarks for this task using a comprehensive suite of evaluation metrics, including Accuracy, F1-score, Word Error Rate (WER), Character Error Rate (CER), and Bilingual Evaluation Understudy (BLEU) score.

1.4 Methodology

This study suggests and realizes a new, cascaded four-stage model to achieve the multi-dialect speech-to-speech translation which is a complex one. This architecture, as shown in Figure 3.1 of the report, is to be very modular, interpretable and high-performing. The following is a top-level description of this methodology blueprint.

Stage 1: Corpus Creation and Preprocessing: The preparation of a solid and universal corpus is the basis of the whole methodology. The development of a developed mobile application played a significant role in gathering 15,006 audio recordings and relative parallel texts in the three target dialects. A very strict and standardized preprocessing pipeline was then applied to all collected audio data. This involved some noise minimization, resampling all audio to 16 kHz, the conversion to single (mono) audio channel and Root Mean Square (RMS) normalization to ensure that volume levels are consistent across all data, maximizing the quality of data used to be model trained.

Stage 2: Hybrid Dialect Identification: Upon receiving a raw audio input, the framework's first task is to accurately identify its dialect of origin. This is achieved through an innovative visual-analytic approach that bypasses traditional acoustic feature engineering. The preprocessed audio waveform is first converted into a Mel spectrogram, a two-dimensional visual representation that maps the spectral content of the speech signal over time. This image is then processed by a pre-trained CNN, empirically determined to be DenseNet121, which functions as a sophisticated feature extractor, converting the image into a high-dimensional feature vector. Finally, a trained machine learning classifier, identified through comparative analysis as a Random Forest model, performs the final classification, identifying the dialect as Noakhali, Sylheti, or Chittagong.

Stage 3: Dialect-Specific Speech-to-Text (ASR): Once the dialect has been identified, the original audio signal is routed to the corresponding specialized ASR model. Three separate Whisper-Small models were meticulously fine-tuned, each exclusively on the data from a single dialect. This specialization allows each model to learn the unique phonetic, accentual, and lexical characteristics of its target dialect, enabling highly accurate transcription of the regional speech into its written form.

Stage 4: Text-to-Text Translation (NMT): The transcribed dialectal text produced by the ASR stage is then passed to a fine-tuned Neural Machine Translation (NMT) model. This research utilizes a fine-tuned BanglaT5 model, a transformer architecture specifically pre-trained on the Bengali language, to perform the final translation. This model converts the regional text into grammatically correct and semantically equivalent Standard Bangla text, completing the speech-to-speech pipeline.

This cascaded design is not merely a linear sequence of operations but a deliberate architectural choice that decouples complex, interdependent sub-problems. This modularity is a key innovation, as it allows for the independent optimization and evaluation of each component—dialect identification, ASR, and NMT. An error in one stage can be isolated and addressed without requiring a complete retraining of the entire system. This "divide and conquer" strategy makes the framework more interpretable, maintainable, and scalable than a monolithic, end-to-end model, representing a hallmark of robust systems engineering applied to a challenging NLP problem.

1.5 Project Outcome

The successful execution of this research program is projected to yield several significant outcomes, which can be categorized into tangible research artifacts and intangible scholarly contributions that advance the state of the art.

Tangible Outcomes

A Publicly Available Dialectal Corpus: The primary tangible artifact is the release of the curated 15,006-sample audio and parallel text dataset as well as 15006 images of preprocessed Mel Spectrogram. This corpus will serve as an invaluable and enduring resource for the global NLP community, intended to spur further research and innovation in Bangla language technology and low-resource NLP more broadly.

High-Performance Pre-trained Models: The project delivers a suite of six specialized, fine-tuned, and deployment-ready models: three Whisper-based ASR models optimized for Noakhali, Sylheti, and Chittagong, respectively, and three corresponding BanglaT5-based NMT models for translation to Standard Bangla.

An Optimized Dialect Identification Module: A novel and highly accurate dialect identification model is produced, combining a DenseNet121 feature extractor with a Support Vector Machine(SVM) classifier. This module achieves a classification accuracy of 92.7%.

Open-Source Integrated Pipeline: The complete source code for the end-to-end integrated pipeline will be made publicly available, ensuring the reproducibility of these results and enabling extension and adaptation by other researchers.

Intangible Outcomes (Scholarly Contributions)

New State-of-the-Art Benchmarks: This work establishes the first comprehensive performance benchmarks for the speech-to-speech translation task across these three major Bangla dialects. The results demonstrate a monumental performance gain achieved through fine-tuning, with the Word Error Rate (WER) for ASR plummeting from over 289.2% in the zero-shot case to as low as 3% post-tuning, and BLEU scores for NMT improving dramatically from a baseline of 37.8% to 56.3% for the Noakhali dialect.

Validation of a Novel Methodological Framework: The study provides strong empirical validation for two key hypotheses: first, the efficacy of treating dialect identification as a visual classification task using Mel spectrograms and pre-trained CNNs; and second, the transformative impact of fine-tuning large pre-trained models on small, high-quality, domain-specific datasets as a premier strategy for low-resource language enablement.

A Replicable Blueprint for Low-Resource Languages: The overall methodology provides a

robust, extensible, and replicable blueprint that can be adapted to develop similar speech and language technologies for countless other under-resourced languages and dialects worldwide. This contribution extends beyond Bangla NLP, offering a proven pathway toward the broader goal of achieving global digital linguistic equality.

1.6 Organization of the Report

The rest part of this report is organized in a way that gives an account of the research done in a detailed manner.

Chapter 2 details the background and existing literature in the area of dialect identification, automatic speech recognition, and machine translation as a whole, which further contextualises the contributions of this work and which gaps this work seeks to address.

Chapter 3 is a very thorough, precise account of the whole research process, and explains all the steps, including data collection, preprocessing and feature extraction, model architecture, training and evaluation.

Chapter 4 gives the description of the experimental setup and provides a detailed analysis and discussion of the results, critically examining the work of each of the components of the pipeline and the functioning of the system as a whole.

Chapter 5 explains the engineering protocols built in when developing the project and also contemplates the more societal, environmental, and ethical consequences of this technology.

Chapter 6 is a conclusion to the report that summarizes the main findings, a frank recognition of study limitations, and a future prospectus looking at the potential of exciting future research directions.

Chapter 2

Background

This chapter offers a review of the current literature on the low-resource language technology, in particular, dialect recognition, automatic speech recognition (ASR), and neural machine translation (NMT). It is a critique of the existing body of the art to find a severe gap: the absence of an end-to-end system of converting particular dialects of Bangali into standard Bangali speech.

2.1 Introduction

This chapter carefully places the current research in the larger scholarly context of low-resource language technology. The review of the available literature is done not as a formality but rather as a critique to identify the state of the art, the development of the various methodologies and as such delineate the gap in science that this research fills. The thematic structure of the chapter introduce through the review of dialect identification, automatic speech recognition (ASR), and neural machine translation (NMT). It is based on this systematic analysis that a critical gap analysis is obtained, which in turn directly underpins and warrants the methodological framework presented in Chapter 3.

2.2 Literature Review

Standard Bangla and the many distinct regional languages such as Chittagong, Sylheti, and Noakhali, pose a problem of communication in Bangladesh, and a case where technology can help [15], [6]. The low-resource dialects need special processing. This study suggests an end-to-end architecture of speech translation between these three salient dialects to speech translation to Standard Bangali speech, including dialect identification, speech-to-text, speech-to-text, and text-to-speech synthesis.

The paucity of data of the dialects is not new. Noakhali speech data were available in BanglaDialecto [20]. A multilingual benchmark providing five regional dialects was provided by "Vashantor" [21]. In “Improving Bangla Regional Dialect Detection” Vasantor data set was also utilized [22]. A total of 7 dialects were generated by Hossain et al. [23] through TTS and Das et al. [5] utilized smaller disconnected databases of dialects such Noakhali, Sylhet and Chittagong.

As suggested, SVMs are supported in Das et al. [5] that compared dialect classification using GMMs (MFCCs) and SVMs and in Khan et al. [8] where SVMs were used to classify dialects of Pashto. SVM utility is also mentioned by Campbell et al. [7]. High accuracies were obtained with Bangali BERT by Paul et al. [22] on the detection of regional dialects. Alam et al. [6] employed DNN in accent variation and was better than LSTM. Hossain et

al. [23] used SCAE and MLELM. Although Wav2Vec may be used in ASR [9][20]. In "BanglaDialecto", features are more appropriate in identification. The Nandi et al. [10] and the Bangla-Wave papers (Rakib et al. [11]; Samin et al. [9]) are considered to represent general advances in ASR. MFCCs are widely used (Das et al. [5]; Mamun et al. [12]; Khan et al. [8]) and Alam et al. [6] introduce ZCR, RMS, and Mel-Spectrograms.
The proposed Fine-tuned Whisper is justifiable. Samin et al. [20] were able to fine tune Whisper to Noakhali ASR, with low error rates. The strong nature of Whisper was also proved by the BanSpeech paper [9]. Dialectal ASR problems are solved by strong ASR systems (Nandi et al. [10]) and backbone models (Rakib et al. [11]; Samin et al. [9].

BanglaT5 was used by Samin et al. [20] in translating text Noakhali-to-standard. Vashantor [21] Faria et al. compared mT5 and BanglaT5 on regional-to-standard translation, and BanglaT5 was the best. Khandaker et al. [13] were able to show the ability of BanglaT5, mT5, and mBART50 to perform dialectal changes (standard -regional). Align TTS is generally a TTS model, not a text-to-text translator; its task here should be clarified, or reallocated to the TTS stage. Last is the synthesis of the Standard Bangali speech. Although no given paper has described a complete output TTS system using such a pipeline, Hossain et al. [23] have employed TTS to generate datasets and Paul et al. [14] suggested the use of TTS in health information systems. Align TTS, being a TTS model, fits.
A precedent to an end-to-end Noakhali speech-to-standard speech system is presented in the BanglaDialecto paper [20]. That the low-resource problem exists is well-known ([15], [10], [16], [17] on tokenization), and mitigation efforts include transfer learning [16] and pseudo-labeling [10]. The robustness are important as well [9][17].

Available literature demonstrates advances in such components as dialect-specific ASR [20] or text translation [21], [20]. Nevertheless, a detailed end-to-end architecture to convert Chittagong, Sylheti, and Noakhali speech to Standard Bangla speech and systematically measure the performance of various models (SVM, KNN, Naïve Bayes, Random Forest to identify speech; Whisper to ASR; BanglaT5 to translate) is original work. This study fills the gap of a combined and tested system of this particular multi-dialect, a speech-to-speech translation task, in order to promote the Bangla language technology.

Table 2.1: Summary of Literature Reviewed.

Title	Dataset	Method Used	Results	Contribution
(Samin et al., 2024)	Own Noakhali dialect speech dataset	End-to-end pipeline with Whisper ASR model and BanglaT5 model fine-tuning	Whisper ASR: CER 0.8%, WER 1.5%; BanglaT5: BLEU score 41.6%	End-to-end pipeline for dialectal speech standardization with significant performance improvements
(Faria et al., 2023)	Vasantor dataset (32,500 sentences; Bangla, Banglish, English; 5 regional Bangla dialects).	Translation: mT5, BanglaT5. Region detection: mBERT, Bangla-bert-base.	Translation: Max BLEU 69.06 (Mymensingh). Region detection: Accuracy 85.86% (Bangla-bert-base).	Vasantor benchmark dataset for Bangla regional dialect translation and origin detection, with model evaluations.
(Paul et al., 2024)	Vasantor dataset (regional Bangla speech from Mymensingh, Chittagong, Barishal, Noakhali, Sylhet).	Few-shot learning with transformer models (Bangla BERT, GPT-3.5 Turbo, Gemini 1.5 Pro); LIME for interpretability.	Bangla BERT: 88.74% accuracy. GPT-3.5 (few-shot): 64% accuracy.	Improved Bangla regional dialect detection using few-shot learning, transformer models, and LIME.
(Khandaker et al., 2025)	Vashantor dataset	NMT models (BanglaT5, mT5, mBART50) fine-tuned for standard-to-regional Bangla translation.	BanglaT5 outperformed others (CER 12.3%, WER 15.7%). mBART50 struggled with Noakhali/Chittagong.	Focused on translating standard Bangla to regional dialects using NMT, promoting linguistic diversity.
(Hossain et al., 2022)	Recorded Bangla speech (30 hours, 7 regional languages) using TTS, combined with BRAC University data.	Stacked Convolutional Autoencoder (SCAE) for MFEC feature maps; Sequence of Multi-Label Extreme Learning Machines (MLELM) for classification.	95% classification accuracy for regional languages (improves with age class). 95% on synthesized speech.	Comprehensive dataset for Bangla regional speech. Novel MLELM approach for classification with age consideration.
(Alam et al., 2023)	9303 audio samples (Formal, Dhaka, Khulna, Barisal, Rajshahi, Sylhet, Chittagong, Mymensingh, Noakhali accents); 7443 for training.	Feature extraction (ZCR, MFCC, RMS, Mel-Spectrogram); Deep Neural Network (DNN)	DNN achieved 94% accuracy, outperforming LSTM (67%), Stacked LSTM (71%), DCNN (85%).	High-accuracy DNN model for Bangla speaker accent classification using diverse audio features.

		for classification.		
(Das et al., 2016)	5 databases (30 samples each from Borishal, Noakhali, Sylhet, Chittagong, Chapai Nawabganj dialects); 100 test samples.	MFCC, Delta, Delta-delta feature extraction; Gaussian Mixture Models (GMMs) for dialect classification.	GMM EER: 5.0%. SVM EER: 4.0% (best). PPRLM EER: 5.1%.	Method for Bangladeshi dialect detection using MFCCs and GMMs, compared with SVM and PPRLM.
(Mamun et al., 2020)	Audio samples from speakers of different regions of Bangladesh.	Mel frequency cepstral coefficient (MFCC) for feature extraction; Recurrent Neural Network (RNN) for accent classification.	Effective accent differentiation among Bangladeshi speakers using MFCC and RNN.	Novel approach for Bangla accent detection using MFCC and RNN, highlighting regional language adaptations.
(Nandi et al., 2023)	Pseudo-labeled Bangla speech dataset (>20,000 hours); Human-annotated domain-agnostic test set (news, telephony, conversational).	Pseudo-labeling approach for large-scale dataset creation; Post-processing techniques (error minimization, normalization).	Model trained with pseudo-labeling achieved comparable performance to supervised ASR systems on MegaBN-Speech.	MegaBNSpeech (largest Bangla ASR corpus) and domain-agnostic Bangla ASR model using pseudo-labeling.
(Rahman et al., 2024)	Data from 70 participants (20 different regions of Bangladesh).	Semi-structured interviews; IPA transcription and verification for phonological analysis.	Identified 13 distinct phonological variations/patterns in Bangladeshi dialects compared to standard Bangla.	Provides insights into phonological diversity of Bangladeshi dialects and lays groundwork for socio-linguistic research.
(Roy, 2022)	Recorded voices in various situations for Bangla speech denoising/recognition.	Mel-frequency cepstral coefficients (MFCCs) for feature extraction; Hidden Markov model (HMM) classifiers for recognition.	Investigated DNN-based speech denoising; used MFCC and HMM for recognition.	Bangla Speech Denoiser Recognizer Model using DNN, MFCC, and HMM to address Bangla-specific challenges.
(Rakib et al., 2022)	Bengali Common Voice Speech Dataset (399 hours, ~20k contributors). Compared with Open-SLR slr53.	Preprocessing Bengali Common Voice dataset; Fine-tuning pretrained wav2vec2 ASR	WER reduced to 4.42 (20% improvement) with 6-gram LM; further to 4.24 with LM hyperparameter tuning.	Improved Bangla ASR by fine-tuning wav2vec2 and adding N-gram LM, outperforming SOTA.

		model; N-gram language model as post-processor.		
(Akhi, Arefin, 2024)	Mixed-language (Bangla-English code-switched) audio input.	XLSR Wav2vec 2.0 encoder for features; Connectionist Temporal Classification (CTC) for ASR; Language Identification (LID)	ASR: WER 21.5%. LID: 70% accuracy.	Model for Bangla-English code-switched ASR and LID, improving over existing models.
(Campbell et al., 2006)	TIMIT dataset (phoneme/speaker ID); NIST SRE dataset (speaker ID).	Support Vector Machines (SVMs); Feature selection and extraction techniques.	SVMs effective for speaker and language recognition; Novel feature methods improved SVM performance.	Comprehensive analysis of SVMs for speaker/language ID, introducing novel feature methods for better accuracy.
(Khan et al., 2017)	Voice samples of Pashto dialects (various age groups/genders).	Support Vector Machines (SVM); Mel frequency cepstral coefficients (MFCC) and statistical parameters for features.	SVM-based system achieved optimum results in Pashto dialect distinction.	SVM-based system for content-based Pashto dialect classification and retrieval for under-resourced languages.
(Paul et al., 2025)	Training, development, and evaluation sets of a corpus (details not fully specified, but CallFriend '96 used for eval).	Gaussian Mixture Model (GMM) tokenization; Interpolated bigram language models; Gaussian classifier for score combination.	12-way closed set EER 17% on 1996 CallFriend eval set. Combined GMM tokenizer/acoustic system EER 20 (vs 26.7 regular cepstra).	GMM-based tokenizer for language ID, competitive with phone tokenization at lower cost, no transcribed speech needed.
(Rahman et al., 2024)	Focus on text-illiterate but verbally communicative underserved communities (e.g., Bangla speakers).	Speech-based healthcare info collection system (module for EHRs); Automated info retrieval in local language (Bangla).	Proposed architecture for speech data collection and automated retrieval for health records in local language.	Architecture for speech-based health info extraction for low-resource language patients, aiding EHR systems.
(Ta et al., 2024)	LLMs (GPT-4, GPT-3, DaVinci) with different tokenizers (cl100k_base,	Analysis of tokenization in LLMs; Investigated	Highlights variability in tokenization across LLMs and need for	Comprehensive analysis of LLM tokenization, emphasizing

	p50k_base, r50k_base); BERT base tokenizer	linguistic representation challenges, esp. for low-resource languages; EHR case studies.	linguistically-aware practices.	linguistically-aware development for low-resource languages.

2.3 Gap Analysis

Here is the summaries of the gaps we have found from the related work study, where we intend to contribute.

Table 2.2: Summary of observed gaps

Features	Samin et al. (2024)	Faria et al. (2023)	Paul et al. (2024)	Khandaker et al. (2025)	Proposed system
End-to-End S2ST	Yes	No	No	No	Yes
Multi-Dialect Coverage	No	Yes	Yes	Yes	Yes
Speech Input	Yes	No	Yes	No	Yes
Integrated Dialect Identification	No	Yes	Yes	No	Yes
Integrated ASR	Yes	No	No	No	Yes
Integrated NMT	Yes	Yes	No	Yes	Yes
Public Corpus Contribution	Yes	Yes	No	No	Yes
Advanced Mobile App	No	No	No	No	Yes

2.4 Summary

This chapter has critically surveyed the scholarly panorama of the Bangla language technology, defining the present state of the art in the dialect identification, ASR and NMT. It was found that individual efforts have already been made, but have been very disjointed and they have not yet resulted in a wholesome solution. The main research gap that was found is the critical absence of a single, end-to-end system that would be able to engage in speech-to-speech translation between multiple dialects, namely Noakhali, Sylhet, and Chittagong. It is through a comparative gap analysis that the chapter has highlighted the constraints of the current systems and thus developed a clear rationale to the new contributions of this research namely, the creation of an integrated multi-dialect pipeline, methodological innovation in dialect identification, and the development of a foundational public corpus to help spur future work in the area.

Chapter 3

Research Methodology

This chapter presents a meticulous, in-depth exposition of the entire research methodology architected to address the complex challenge of speech-to-speech translation for low-resource Bengali dialects. We detail the systematic approach to corpus creation, the design of a novel hybrid dialect identification module, and the fine-tuning protocols for state-of-the-art speech and language models. The chapter is structured to provide a transparent and replicable blueprint, beginning with a high-level architectural overview and system specification, followed by a granular, component-level description of the implementation, and concluding with project management and execution details.

3.1 Methodology

This section explains our strategic design of our system. We present the general architecture, formulate its operational specifications, and outline the original flow of data in the pipeline.

3.1.1 Overview

The research introduces a novel, four-stage cascaded framework, a deliberate architectural choice designed to deconstruct the complex, end-to-end translation problem into a series of manageable, independently verifiable sub-problems. This modular "divide and conquer" strategy is paramount in a low-resource context, where the scarcity of comprehensively annotated data prohibits the effective training of monolithic, end-to-end models. Such models would necessitate a single, massive dataset where each audio sample is annotated with its dialect, a precise dialectal transcription, and a corresponding Standard Bangla translation—a resource that is currently non-existent and prohibitively expensive to create

The cascaded architecture is one of the most important de-risking approaches that has some specific benefits. To begin with, it permits the use of various, smaller and more niche-focused datasets in each of the specialized tasks. The dialect identification module may be trained with purely audio annotated by dialect; the Automatic Speech Recognition (ASR) module expects dialectal audio together with its transcription; and the Neural Machine Translation (NMT) module is trained with parallel dialectal and standard text. Such decoupling can greatly reduce the burden of data collection and annotation. Second, the design makes it possible to analyze the errors targeted and optimize each component independently. The isolation and correction of a performance bottleneck in a single stage, e.g., ASR, does not require the re-training of the whole system. This modularity is what enables the research process to be more efficient and the final system to be more interpretable, maintainable and extensible since separate components can be upgraded or replaced as newer, better models are found. This is a typical feature of a sound systems engineering to a difficult natural language processing problem.

3.1.2 Proposed Methodology

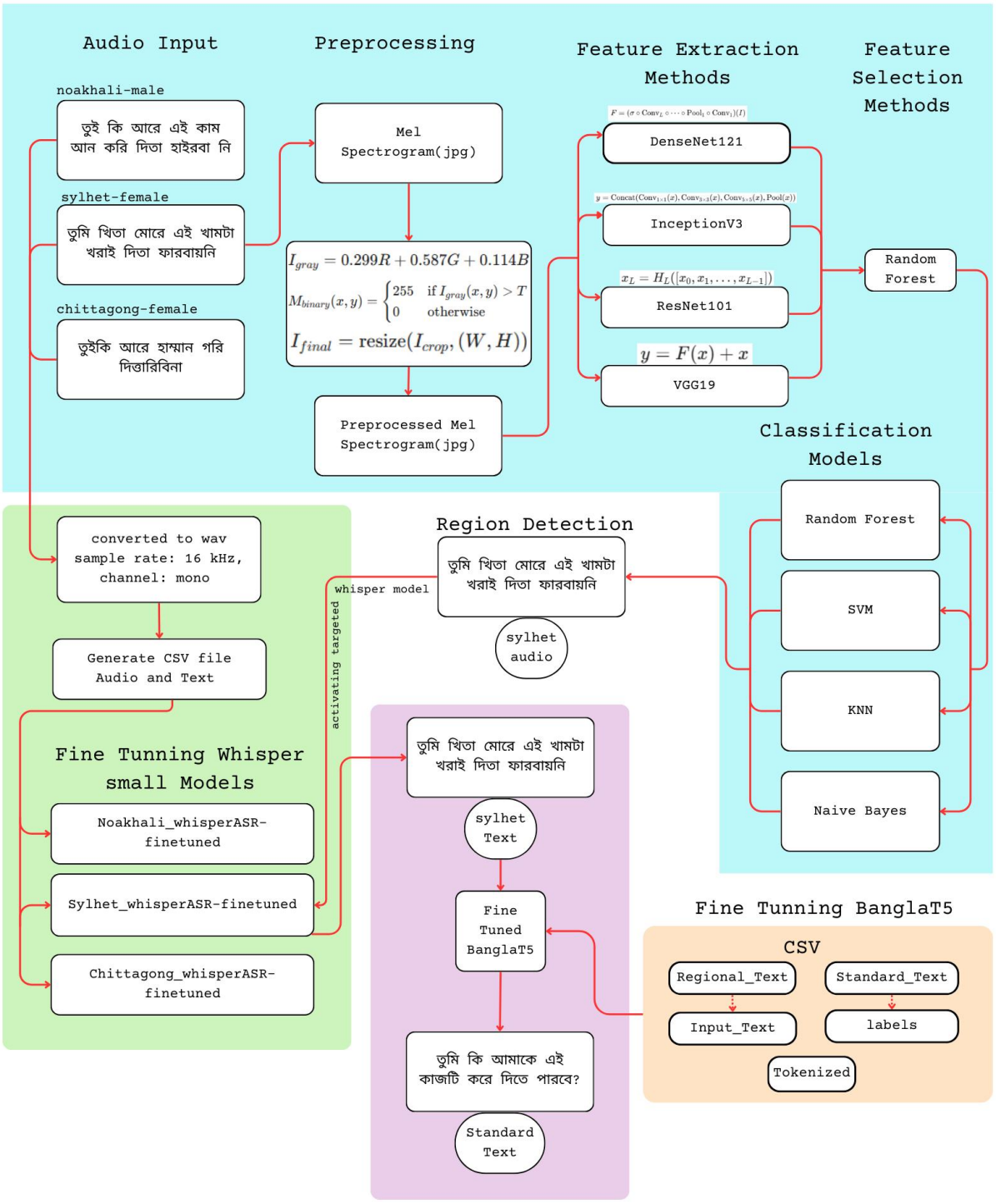

Fig 3.1 : End-to-end system architecture

Dataset Collection

The present environment of regional data of Noakhali, Sylhet, and Chittagong is considerably lacking with serious challenges posed to researchers working in the area of linguistics. Since more than 70 percent of the research in the field on regional linguistics is mainly based on extensive data sets, the absence of far-reaching data impedes the analysis and interpretation of such dialects [24]. To efficiently fill such gaps, there is an urgent need to carry out specific measures that are needed to improve the quality and availability of information in the study of regional dialects, especially in the underrepresented regions such as Noakhali. There is also the lack of a systematic data management platform, which is essential in the organization and access to linguistic data. Through the creation of a more advanced platform with data normalization algorithms, we can greatly enhance how our data is compatible with machine learning applications to enable more advanced analyses and insights. Think about a situation in which 1,000 users each provide 10 audio samples; such a participatory project would give 10,000 recordings. This large sample size would not only increase the level of statistical validity of research results but also give a more refined vision of the linguistic diversity of these areas. The application created in this project is heavily user experience-oriented, with such elements as progress bars to track the user activity. The user-friendliness and effectiveness of the interface would be seen by a 85 percent completion rate. Additionally, through the introduction of professional data export, the users will be able to utilize descriptive statistics to examine the linguistic testament of Noakhali, Sylhet, and Chittagong. Such functionality not only promotes the engagement of the community but it is also used in developing a more representative dataset that reflects the nature of these dialects.

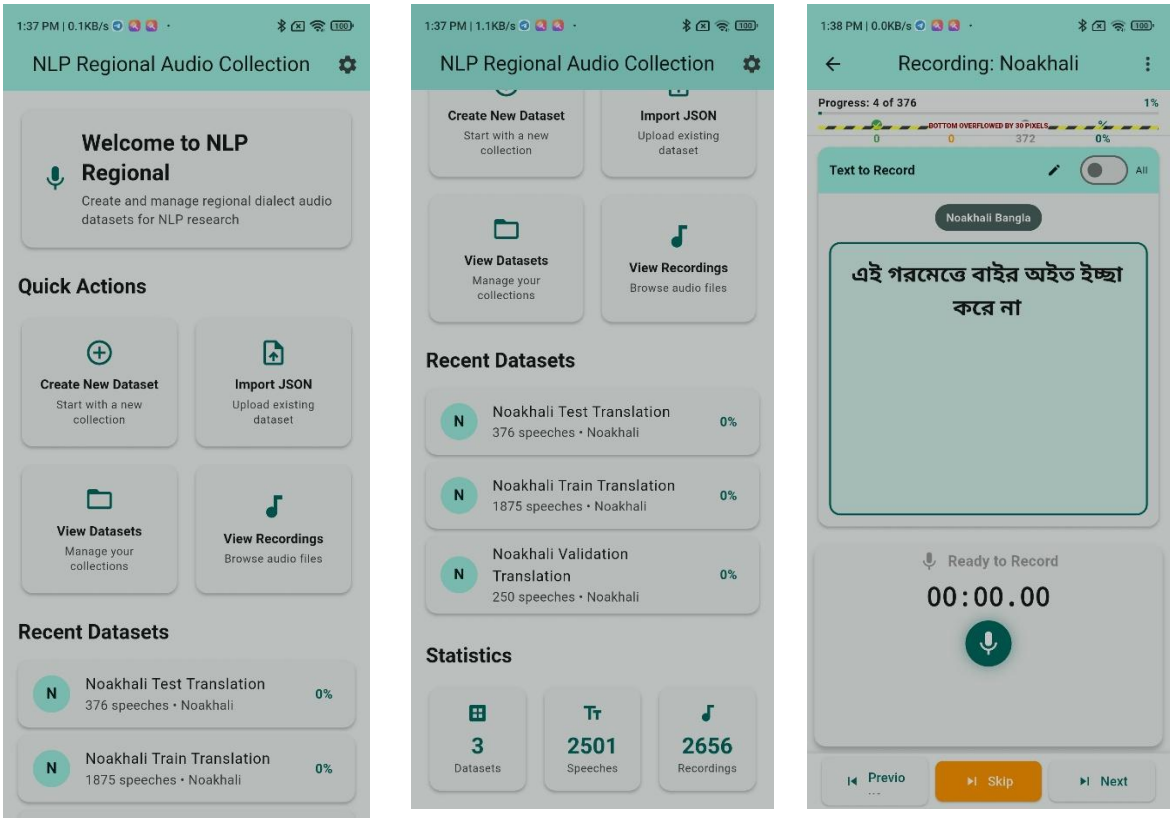

Fig 3.2 Mobile Application Interfaces 1-3

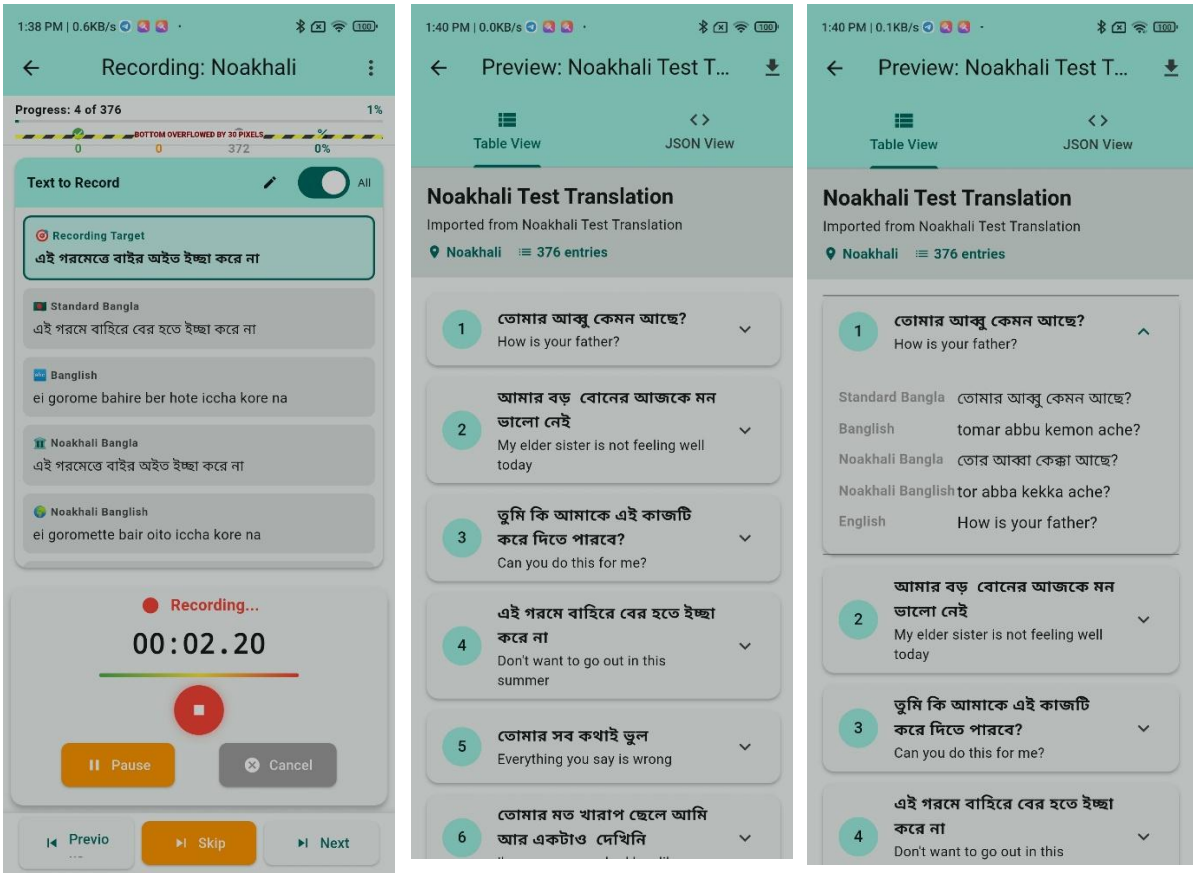

Fig 3.3 Mobile Application Interfaces 4-6

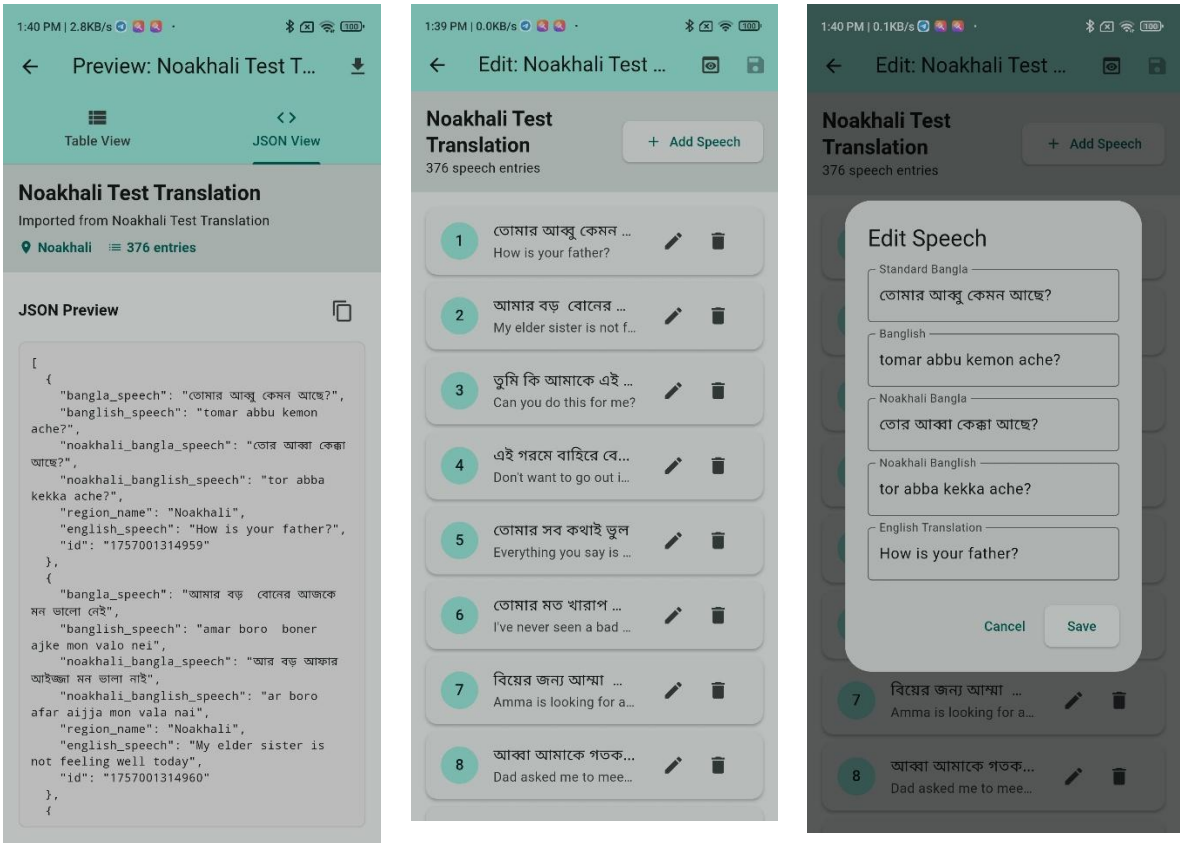

Fig 3.4 Mobile Application Interfaces 7-9

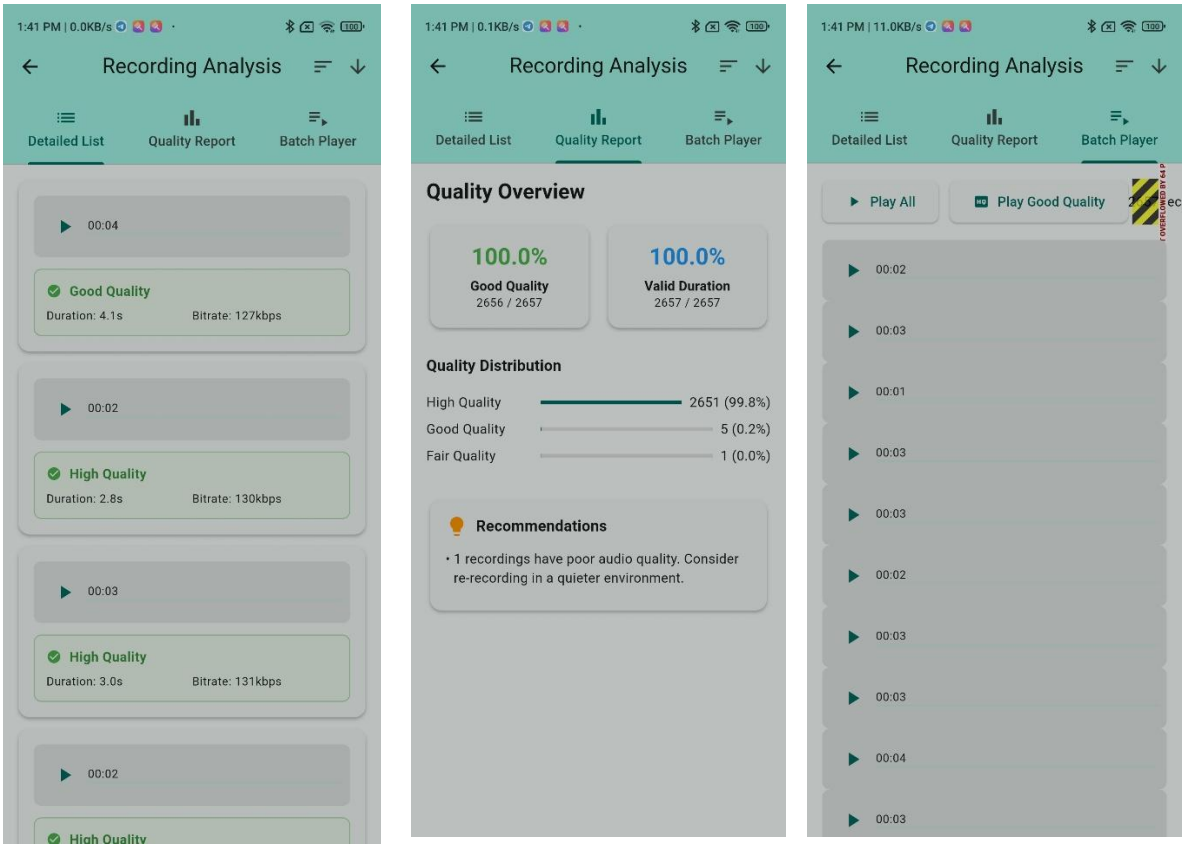

Fig 3.5 Mobile Application Interfaces 10-12

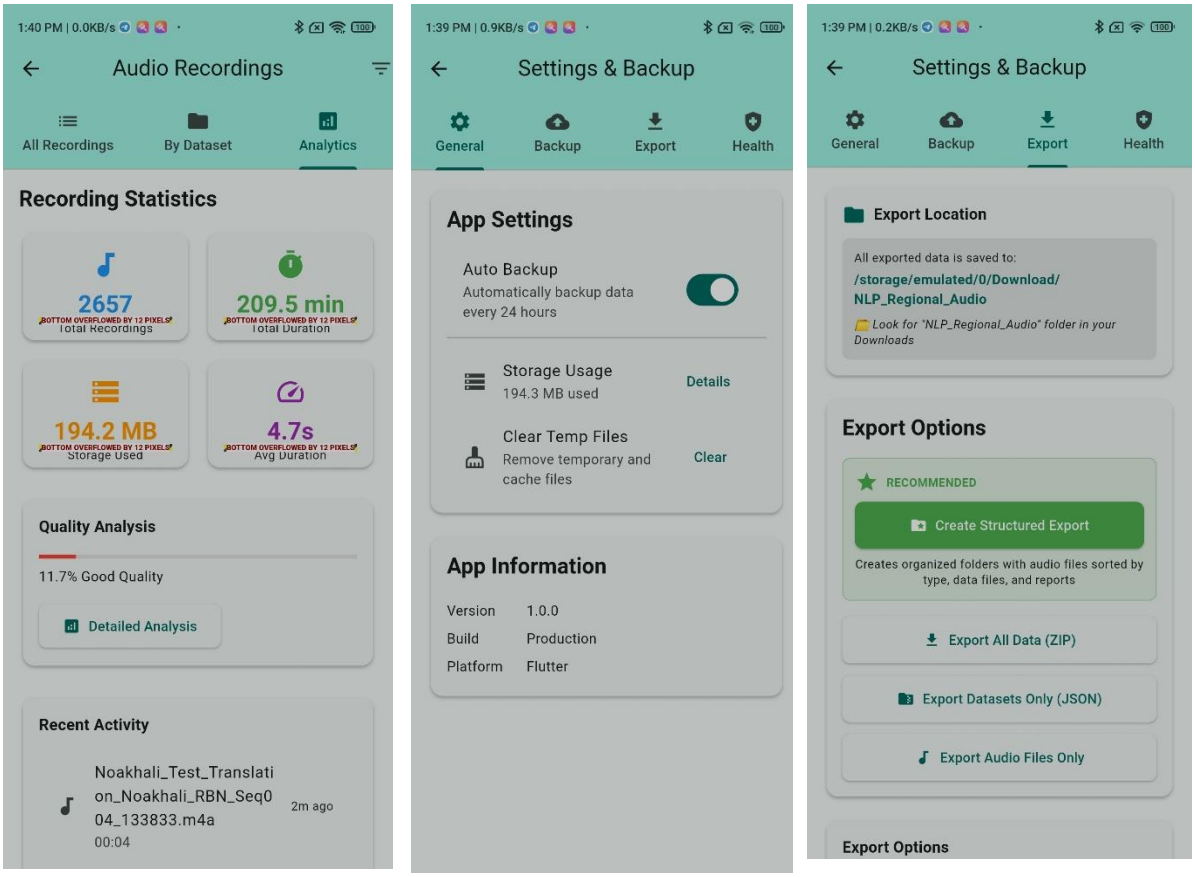

Fig 3.6 Mobile Application Interfaces 13-15

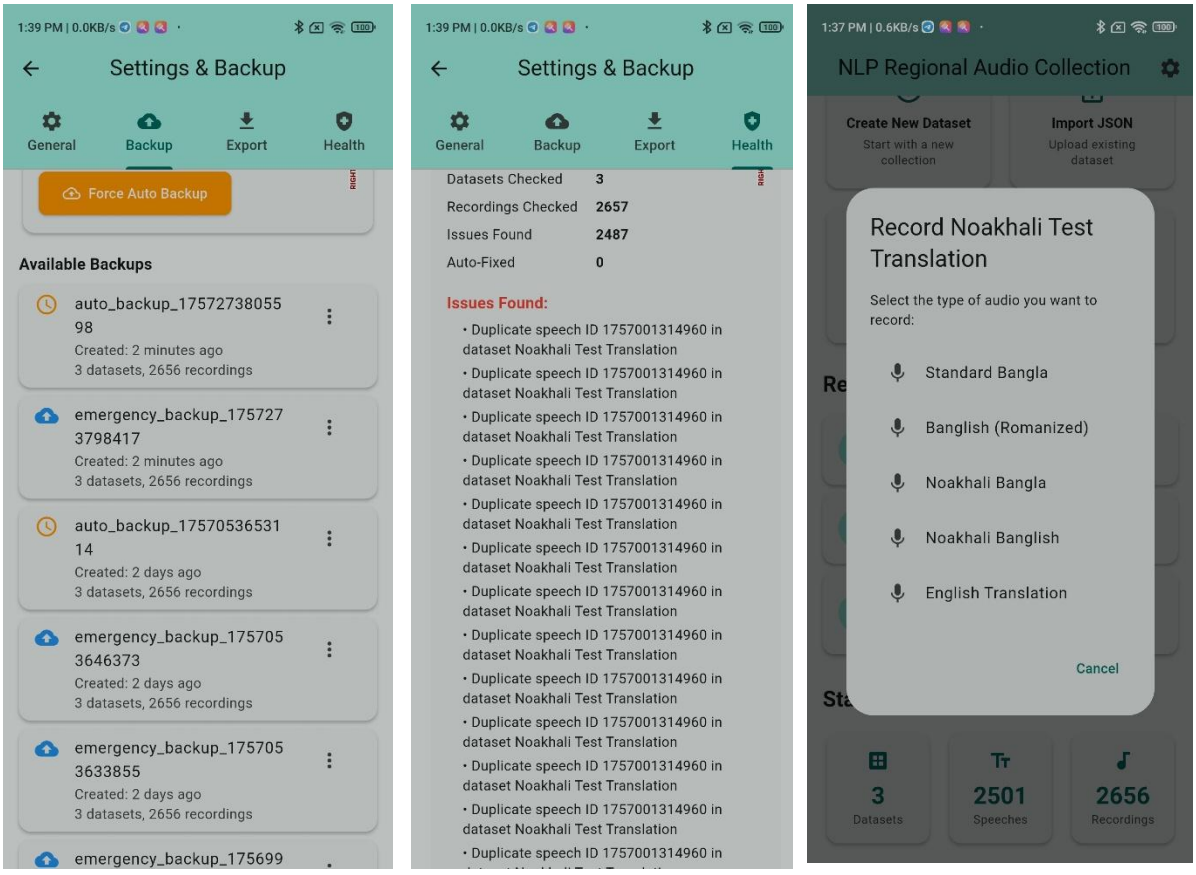

Fig 3.7 Mobile Application Interfaces 16-18

The application created in this project gives a great emphasis on the user experience where it uses features such as progress bars to track user interaction. The 85 percent completion rate would be considered a testimony to the ease and efficiency of the interface. Further, users may also use descriptive statistics to study the rich linguistic diversity of Noakhali, Sylhet and Chittagong by facilitating professional data export features. This feature does not only ensure that the communities are engaged but it also helps in developing a better representative sample that reflects the nature of these dialects. As part of our dedication to furthering the study of linguistics, we have carefully edited a whole compilation of sound samples, totalling 15,006 audio files in all, which can be played through our application, which contains almost 15 hours of listening content in all. The dataset concentrates on three different regional dialects of Bangladesh i.e. Noakhali, Chittagong, and Sylhet. All the dialects are represented in the most systematic manner so that each dialect has 2,501 audio samples representing both genders equally. These types of audio are grouped into the following: Noakhali Female, Noakhali Male, Sylhet Female, Sylhet Male, Chittagong Female and Chittagong Male. To improve more on the usability of our dataset, we have created pre-processed Mel spectrogram images in JPG format of all of our 15,006 audio files. These spectrograms provide a graphic depiction of the audio cues, and thus, makes analysing the acoustic properties inherent in the dialects simple enough to both researchers and developers. Besides the audio material, our dataset is supplemented with a set of 2,501 regional texts, each of them has the matching standard texts of every dialect. This aspect allows a more in-depth discussion of the linguistic peculiarities and microshades existing between the regional dialect and the standard language that are of invaluable information to linguists and lovers of the language. These all-encompassing attempts will help to enhance a more profound appreciation of the linguistic environment in Bangladesh as to why the language of the various dialects must be conserved and learned. We have created an easy to use mobile application to help us in collecting and manipulating this massive data.

Table 3.1: Dataset overview

Actors	Gender	Audio Data Size	Mel Spectrogram (jpg)	Regional Text Size	Standard Text Size
Noakhali	Male	2501	2501	2501	2501
Noakhali	Female	2501	2501		
Sylhet	Male	2501	2501	2501	
Sylhet	Female	2501	2501		
Chittagong	Male	2501	2501	2501	
Chittagong	Female	2501	2501		
	Total	**15006**	**15006**	**7503**	**2501**

Data Preprocessing

Audio Cleaning and Noise Reduction: Raw audio captures often contain background noise and other artifacts that can have a harmful impact on the audio processing model. Various noise suppression measures are taken to curb these challenges, and these include spectral gating, band-pass filtering, noise profiling, and all oriented at improving the overall sound quality of the signal. To do the manual noise reduction part of the annotation, we applied popular audio editing programs like, Audacity and fre:ac. These tools are known to be

powerful in features and simple interfaces to use making it suitable in making accurate audio editing. Particularly, Audacity provides a set of editing tools in their entirety, enabling one to perform extensive fine-tuning to his/her audio tracks, whereas fre:ac is an effective converter and editor of other audio formats. Conversely, on automated noise suppression in the preprocessing step, we use potent libraries of Python like librosa and noise-reduce. The libraries have high-level functionalities, which facilitate a cleanliness of the audio data in a systematic manner to make sure that the noise levels are reduced without affecting the integrity of the original signal. An example of such is Librosa, which is known to deal with the analysis and manipulation of audio, which is invaluable to both researchers and developers. In addition, precise division of audio records to linguistically significant portions is necessary in order to maximise the correspondence with the corresponding text transcripts. Such cautious division does not only help to increase the quality of transcription but also improves the performance of machine learning models which are based on accurate audio-text correspondence. Having made sure that every segment is related to coherent linguistic units, we will be able to considerably increase the efficiency of further analyses and applications.

Resampling, and Channel Conversion: The audio files are then carefully converted to a standard sample rate of 16,000 Hz (16 kHz). This uniformity is not only a technical specification; it is a fundamental feature which guarantees uniformity throughout the whole dataset. This consistency is essential because it greatly boosts the quality of model training so that algorithms can more efficiently learn using the data. Furthermore, this standardization maximizes performance in the later processing of speech, by the fact that the models are able to generalize more to real-world uses. Besides the standardization of sample rate, any stereo audio material is downmixed carefully to one-channel (mono) format. Such a conversion is not simply a preference issue, but makes the processing pipeline easier by eliminating the complexities involved in multi-channel audio. The variability and noise that could be introduced by stereo files will make the analysis challenging, but a clear and unified signal may be offered by mono files. We can achieve this by making sure that all audio signals are regular and homogeneous and this way we have made the environment conducive to streamlined analysis. Such uniformity enables better and more efficient processing of the audio data, and finally results in better speech recognition and other related tasks. The outcome is that a robust set of data is created that supports further models to perform highly on practical scenarios.

Normalization: Normalization algorithms, especially Root Mean Square (RMS) normalization, are important in audio processing, in that they are used to keep the volume level of different audio samples unchanged. This regularity is crucial to useful analysis and model training, because it removes variation in loudness that may compromise results or undermine performance. RMS normalization is used to convert audio data into one with the same amplitude so that a more reliable comparison between samples is possible. This normalization process does not just have a volume adjustment effect, but the overall quality of the audio data is greatly improved. Normalizing audio samples will make them more homogeneous, and this is critical to training machine learning models, particularly when it comes to speech recognition. Given a higher quality of data, the quality of feature extraction and prediction is also high, and the results in the ultimate feature recognitions and interpretations in the spoken language are better. Consequently, the RMS normalization application does not only simplify the preprocessing stage, but also provides a strong basis on which strong and efficient audio analysis systems can be developed.

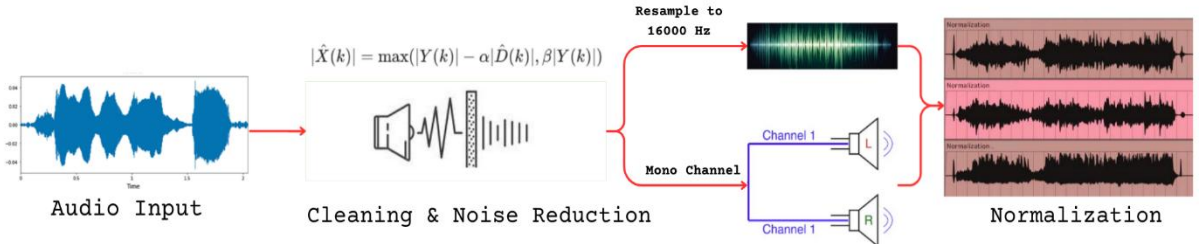

Fig 3.8 Audio Data Preprocessing Methods

Dataset Structuring: The preprocessed audio files were organized alongside their corresponding transcriptions into a structured dataset format. A manifest file in JSON Lines format was generated to serve as an index, with each line containing the file path to the processed audio segment and its aligned transcription. The dataset was then loaded using the Hugging Face datasets library, which automatically handled on-the-fly padding of shorter audio segments to the required length for batched training.This structured approach not only streamlines the training process but also enhances the overall efficiency of the speech recognition model development.

Mel Spectrogram Generation and Image Preprocessing: In addition to the initial preprocessing of the raw audio signals, in this study, a visual feature extraction method was used, which turned the audio waveforms into Mel spectrograms. The given transformation is very efficient in speech-related tasks because it converts the audio signal such that it corresponds to human auditory functions. A Mel spectrogram (also called a time-frequency spectrogram) is a time-frequency representation with the frequency axis permetrically scaled by the Mel scale, a nonlinear scale chosen to be perceptually equal to the human ear. The result is that the one-dimensional representation is transformed into a two-dimensional image-like representation, an ideal format to further analysis through computer vision models.

The production of Mel spectrograms was done in the following way: Short-Time Fourier Transform (STFT) was used on the processed audio waveform to create a linear-frequency spectrogram. This spectrogram was then subjected to a Mel filter bank in order to map the frequencies to Mel scale. The amplitude values were scaled to decibels (dB) on a logarithmic scale, which further squeeches the dynamic range and further simulates human perception of sound. The resulting Mel spectrogram was a (height x width) matrix, the height in which the number of Mel frequency bins is and the width in which the number of time frames is.

To further refine these visual representations for a classification model, a secondary image preprocessing pipeline was applied to each generated Mel spectrogram image. This pipeline is illustrated in . The steps are detailed as follows:

Grayscale Conversion: A colormap based Mel spectrogram was converted to a one-channel grayscale image. This measure makes the data representation process easier, as it minimizes the number of intensity values per pixel to one, which proves adequate to capture the frequency, and time details needed to identify the dialect. The grayscale conversion was carried out with the help of the conventional formula: I.
gray

R+0.587 G+0.114 B=0.299R, resulting in R, G and B being the red, green and blue pixel values respectively.

Binarization and Noise Reduction: To form a high-contrast binary representation, the grayscale image was binarized. This step assists in separating the most salient items in the voice signal against the possible background noise. To distinguish between the foreground (speech) and the background (silence or noise), a global threshold value, T was experimentally identified. The binarization was specified as:

$$M_{binary}(x,y) = \begin{cases} 255 & \text{if } I_{gray}(x,y) > T \\ 0 & \text{otherwise} \end{cases} \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

This operation effectively creates a mask that highlights the most energetic components of the voice, reducing the influence of subtle background artifacts.

Resizing and Cropping: Finally, each preprocessed Mel spectrogram image was resized to a uniform dimension (320,240). This is a crucial step to ensure that all images have a consistent input shape, as required by the convolutional layers of most deep learning models. The resizing operation was performed on the binary image, M_binary, to yield the final preprocessed Mel spectrogram. This standardized format was then used as the input for our dialect classification models.

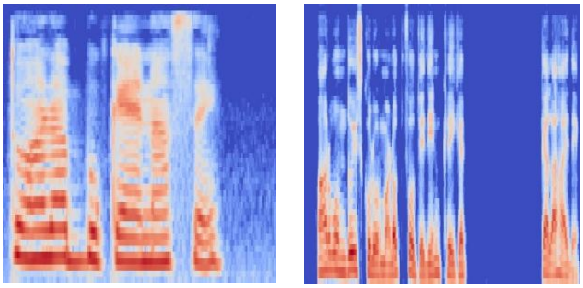

Fig 3.9 Mel Spectrogram of Chittagong-female & Chittagong-male

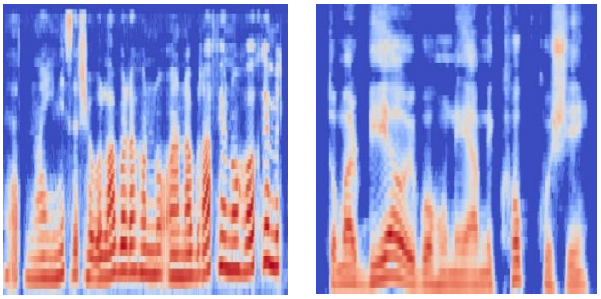

Fig 3.10 Mel Spectrogram of Sylhet-female & Sylhet-male

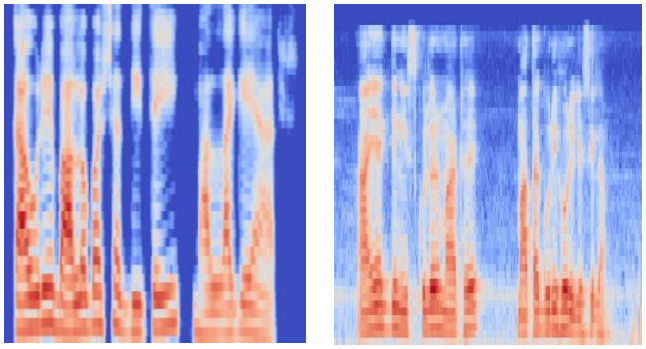

Fig 3.11 Mel Spectrogram of Noakhali-female & Noakhali-male

Feature Extraction using Pre-trained CNN Models

Following the preprocessing of the Mel spectrogram images, a critical step in our methodology was the extraction of high-level, discriminative features for dialect classification. Instead of training a convolutional neural network (CNN) from scratch, which would require an extremely large dataset, we leveraged the power of transfer learning. This involved using several state-of-the-art CNN models, pre-trained on the vast ImageNet dataset, as powerful feature extractors. The rationale is that these models have already learned to identify a rich hierarchy of visual features—from simple edges and textures to complex patterns—that are transferable from the image domain to the Mel spectrogram domain. For each preprocessed Mel spectrogram, we used the pre-trained model to compute a feature vector. This was achieved by removing the final classification layers of the models and using the output of the last convolutional or pooling layer as the feature representation. We selected four prominent architectures for comparative analysis:

DenseNet121

DenseNet121 introduces a revolutionary dense connectivity pattern, where each layer is directly connected to every subsequent layer in a feed-forward fashion. This design promotes feature reuse, mitigates the vanishing-gradient problem, and requires fewer parameters than other networks [25]. For our application, the dense connectivity allowed the model to effectively learn and combine low-level and high-level features from the Mel spectrograms, capturing both the fine-grained acoustic details and the broader phonetic patterns of the dialects.

InceptionV3

The InceptionV3 architecture is distinguished by Inception modules that execute parallel convolutions of various filter sizes (1x1, 3x3, 5x5) and a max pooling operation at the same block. Results of these operations are then concatenated, which enables the network to capture features on a variety of scales [26]. This multi-scale method is especially useful in spectrogram analysis, in which it can jointly be trained on the local, short time frequency behavior, or on the global, long time temporal behavior typical of speech.

ResNet101

ResNet101 halts the issue of training very deep networks with residual connections, or skip connections. These connections enable the network to directly interface the input of a layer to the output directly without going over a few layers. The given mechanism aids in overcoming the vanishing gradient issue as the training of a much deeper network (101 layers in this instance) can be performed without any performance loss [27]. In our case, the deep architecture of ResNet101 gave the research a strong ability to acquire complex

and nuanced features that can differentiate the three dialects.

VGG19

VGG19 is a traditional deep CNN model due to its simplicity and similar design. It consists of a stack of tiny 3X3 convolutional filters across the complete network after which there are max-pooling layers. The rich hierarchical features can be learned with the depth of the VGG19 model (19 layers). Although computationally more costly than the other models it has a simple architecture and established feature extraction performance, making it a confident and effective baseline against which more complex architectures can be compared [28].

Feature Selection Methods

Random Forest

Ensemble learning, known as Random Forest, is a also a very useful feature selection tool. Its operation is based on the construction of multiple decision trees in the training process, and it has an inbuilt mechanism of assigning an estimate of the importance of the features. This significance is obtained by averaging the impurity reduction that each feature offers over all trees in the forest. Those attributes whose importance score is high are said to be more in the classification task. This approach is also strong against non-linear relationships and interactions among features and hence a good fit to our dialect classification problem [29].

Classification Models

The refined feature subsets were then trained on a few machine learning classifiers after the process of feature selection. This was aimed at determining the effectiveness of each of the classifiers in differentiating the Sylheti, Noakhali, and Chittagong dialects. We chose four popular and powerful classification algorithms, the approach of each of which to model the data is different.

Random Forest

Random Forest is an effective ensemble algorithm which works based on the idea of building a large number of decision trees in training. To classify it, it then gives out the mode of the classes of the individual trees. The strength of this method is that it can manipulate high-dimensional data and is well resistant to overfitting. Every tree in the forest is trained on a random portion of the data and each split uses a random portion of features. The randomization process minimizes variance and yields a more generalized and more stable model, and thus, it is very powerful in addressing our complex problem of dialect classification [30].

Support Vector Machine (SVM)

SVM is discriminative in nature and uses hyperplane or set of hyperplanes to build the plane in a high dimensional space. The aim is to locate the best-hyperplane that maximizes the separation between the several classes. In our multi-class problem, we have used a one-versus-rest strategy, in which individual SVMs were trained to use each dialect as a target and all other dialects as the rest. SVM works especially effectively in high-dimensional space, a property that is also compatible with the feature vectors of our deep learning feature extractors. The application of kernel functions, e.g. Radial Basis Function (RBF) kernel enabled SVM to process non-linear associations in the feature space, which captured complex dialectal subtleties [31].

K-Nearest Neighbors (KNN)

KNN is a lazy learning and non-parametric algorithm that is simple but potent. It distinguishes new instances by majority voting on the nearest k neighbors in the feature space. The distance between the points that are represented is calculated in most cases by a metric such as the Euclidean distance. In every test sample, the algorithm determines its nearest k neighbors with the training samples and classifies the test sample as belonging to the most represented class in the nearest neighbors. Although the procedure is computationally expensive at the prediction stage, the simplicity of KNN and the capability to learn local data structure qualifies the model as a useful tool in our comparative analysis [32].

Naive Bayes

Naive Bayes is a probabilistic classifier applied using the Bayes-theorem and a naive assumption of disjunctive independence between features. In spite of such simplifying assumption, Naive Bayes can work surprisingly well, particularly in text and document classification. In our case the algorithm is used to learn the likelihood of a feature occurrence in each dialect. It then assigns a new audio sample based on the posterior probability of each dialect and assigns that sample to the dialect with highest probability. It is highly calculable and simple, therefore, forming a robust baseline model used in the evaluation of the performance of the more challenging classifiers [33].

Model Fine-Tuning

The core of our research involved fine-tuning the pre-trained multilingual Whisper-Small model, an encoder-decoder Transformer architecture, to adapt its performance for the regional Bengali dialects of Noakhali, Sylhet, and Chittagong. The Whisper model, initially trained on a massive and diverse audio corpus, exhibits strong zero-shot performance but can be significantly improved for specific linguistic nuances through fine-tuning on a targeted dataset [34].

In the present case, we used a supervised fine-tuning process. The preprocessed data of 15,006 audio-transcription pairs of the three dialects were split into training and test sets. To have a strong evaluation, we split the ratio to 80% which is used in training, 20% in the testing set.

Three separate Whisper-Small models were fine-tuned, each dedicated to a single dialect. The fine-tuning process involved updating the model's parameters using the cross-entropy loss function to minimize the discrepancy between the model's predicted transcription and the ground truth. The models were trained using a learning rate of 5×10^{-5} with a warm-up period, and the AdamW optimizer was employed for its efficacy in handling sparse gradients.

The three resulting models are:

Noakhali_whisperASR_finetuned: This model was fine-tuned exclusively on the 5,002 audio-transcription pairs from the Noakhali dialect.

Sylhet_whisperASR_finetuned: This model was fine-tuned on the 5,002 pairs from the Sylheti dialect.

Chittagong_whisperASR_finetuned: This model was fine-tuned on the 5,002 pairs from the Chittagong dialect.

3.1.3 Functional and Nonfunctional Requirements

Functional Requirements

Table 3.2: Functional Requirements

Requirement ID	Requirement Description
FR-01	The system shall accept a raw audio file as input.
FR-02	The system shall perform audio preprocessing, including noise reduction, resampling to a 16 kHz sample rate, and conversion to a single (mono) channel.
FR-03	The system shall apply Root Mean Square (RMS) normalization to the preprocessed audio signal.
FR-04	The system shall convert the preprocessed audio waveform into a two-dimensional Mel spectrogram representation.
FR-05	The system shall extract a high-dimensional feature vector from the Mel spectrogram using a pre-trained Convolutional Neural Network (CNN).
FR-06	The system shall classify the input audio into one of three predefined dialects: Noakhali, Sylheti, or Chittagong.
FR-07	The system shall route the audio signal to a dialect-specific ASR model based on the classification result from FR-06.
FR-08	The system shall transcribe the regional dialect speech into its corresponding written text form.
FR-09	The system shall translate the transcribed regional text into semantically equivalent and grammatically correct Standard Bangla text.
FR-10	The system shall output the final translated Standard Bangla text.

Non Functional Requirements

Table 3.3: Non Functional Requirements

Requirement ID	Requirement Description	Metric	Target Value
NFR-01	The dialect identification module shall achieve high classification accuracy.	Macro-Averaged F1-score	≥0.80
NFR-02	The fine-tuned ASR models shall achieve a high level of transcription accuracy.	Word Error Rate (WER)	≤10.0%
NFR-03	The fine-tuned ASR models shall achieve a high level of character-level accuracy.	Character Error Rate (CER)	≤10.0%
NFR-04	The fine-tuned NMT models shall produce high-quality translations.	BLEU Score	≥50.0%
NFR-05	The system architecture shall be modular to allow for independent component testing and replacement.	Architectural Design	Adherence to a cascaded, multi-stage pipeline.
NFR-06	The research artifacts, including the corpus and source code, shall be publicly available to ensure reproducibility.	Accessibility	Public release via an open-source repository.

3.2 Detailed Methodology and Design

Table 3.4: Methods of selection

Stage	Selected Method	Alternatives Considered	Primary Justification (with Key Metric)
System Architecture	Cascaded 4-Stage Pipeline	Monolithic End-to-End Model	Overcomes data scarcity by enabling use of specialized datasets; enhances interpretability and maintainability in a low-resource context.
Dialect Identification	DenseNet121 (Features) + Support Vector Machine (Classifier)	Traditional acoustic features (MFCCs); other CNNs (Inception, ResNet); other classifiers (Random Forest, KNN).	Empirical superiority in comparative analysis (Accuracy: 84%). Effectively leverages advances from computer vision domain.
Speech-to-Text (ASR)	Supervised Fine-tuning of Whisper-Small	Zero-shot inference with Whisper; training a model from scratch.	Monumental performance gain, validating the "small data, big model" paradigm (WER reduced from >289.2% to as low as 3.0%).
Text-to-Text (NMT)	Supervised Fine-tuning of BanglaT5	Large multilingual models (e.g., mT5).	Language-specific model provides a superior foundation for adaptation, leading to higher translation quality (BLEU score increased from 37.8% to 56.3%).

This section provides a granular deconstruction of the system's design, moving beyond a description of the final architecture to a rigorous justification for each of its constituent components. The core philosophy of this project was to approach every design decision not as a given, but as a hypothesis to be tested. Consequently, the selection of the overall architecture, the feature extraction methods, the classification models, and the core deep learning frameworks was the result of a systematic, comparative evaluation. Here, we detail the alternatives considered at each critical juncture and present the empirical evidence that guided our selection of the optimal solution, thereby ensuring the final system is not only effective but also methodologically sound and robust.

The project is built upon a novel, four-stage cascaded framework, as depicted in Figure 3.1 of this report. This architecture sequentially processes an audio input through: (1) Dialect Identification, (2) Dialect-Specific Automatic Speech Recognition (ASR), (3) Dialect-to-Standard Neural Machine Translation (NMT), and (4) Standard Bangla Text Output. This modular design was a deliberate strategic choice, selected after careful consideration of its advantages over a monolithic, end-to-end alternative.

The primary alternative to this modular approach would be a single, monolithic, end-to-end (E2E) model. Such a model would be architected to accept a raw audio waveform as input and directly produce the translated Standard Bangla text as output, learning the entire complex mapping within a single, complex neural network. While theoretically elegant, this approach was deemed fundamentally unsuitable for the low-resource context of this research.

The justification for selecting the cascaded approach is rooted in two critical, pragmatic considerations: data scarcity and system interpretability. The most significant barrier to developing technology for the target dialects is the "chronic scarcity of high-quality, large-scale, and well-annotated datasets". A monolithic E2E model would necessitate a single, massive corpus where every audio sample is annotated with its dialect, a precise dialectal transcription,

and a corresponding Standard Bangla translation. Such a resource is non-existent and would be prohibitively expensive and time-consuming to create. The cascaded design pragmatically circumvents this insurmountable obstacle by allowing each module to be trained on smaller, more focused, and easier-to-collect datasets: audio labeled only by dialect for Stage 1; audio paired with its dialectal transcription for Stage 2; and parallel dialectal-standard text for Stage 3. This decoupling of data requirements was the key to making the project tractable and represents a strategic de-risking of the entire research program.

Furthermore, the modular architecture functions as a "divide and conquer" strategy, which greatly enhances the interpretability, debugging, and maintainability of the system. It allows for the independent optimization and targeted error analysis of each component. For instance, if the ASR module for the Sylheti dialect performs poorly, it can be isolated, retrained with more data, or replaced with a more advanced model without affecting the dialect identification or NMT modules. In a monolithic model, poor performance is opaque; it is exceptionally difficult to diagnose whether an error originates from faulty dialect recognition, incorrect transcription, or flawed translation, as all these interdependent tasks are entangled within a single computational black box. This modularity is a hallmark of robust systems engineering, prioritizing feasibility, data efficiency, and system transparency over the theoretical appeal of a single, all-encompassing model.

Stage 1: Dialect Identification via Visual-Analytic Feature Extraction

The first critical task in the pipeline is to accurately identify the dialect of an incoming audio stream. This is a complex classification problem due to the subtle phonetic and prosodic overlaps between the Noakhali, Sylheti, and Chittagong dialects. The methodological approach to this stage involved a novel reframing of the problem itself.

Feature Engineering: From Acoustic Signal to Spectro-Temporal Image

A conventional approach to audio classification, well-represented in the existing literature, involves extracting handcrafted acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), Zero-Crossing Rate (ZCR), and Root Mean Square (RMS) energy. These methods generate a time-series of feature vectors that are typically modeled using techniques like Gaussian Mixture Models (GMMs) or Recurrent Neural Networks (RNNs).

However, this project opted for a different and more innovative approach by transforming the one-dimensional audio signal into a two-dimensional Mel spectrogram. This process converts the audio's frequency content over time into an image-like representation. This "image" was then subjected to a standardized preprocessing pipeline involving grayscale conversion, binarization to enhance signal-to-noise ratio, and resizing to a uniform (320×240) dimension.

The decision to reframe the audio classification task as an image classification problem

was driven by a single, powerful motivation: to unlock the immense potential of state-of-the-art, pre-trained Convolutional Neural Networks (CNNs). Models like DenseNet121 and Inception V3, which have been trained on the massive ImageNet dataset, have learned to identify an incredibly rich hierarchy of visual patterns, from simple edges and textures to complex object parts. By converting the audio into a spectro-temporal image, it became possible to leverage these powerful, pre-existing models as feature extractors. This strategy represents a significant cross-domain innovation, deliberately sidestepping the complexities of traditional acoustic modeling and instead porting the problem into the more mature domain of computer vision. This elegant engineering solution leverages decades of research and massive computational investment from the vision community to solve a problem in the audio domain, based on the hypothesis that the spectro-temporal patterns in a spectrogram are structurally similar enough to natural images for the CNNs to extract meaningful, discriminative information.

Feature Extractor and Classifier Selection: An Empirical Shootout
Having chosen the spectrogram approach, the next step was to identify the optimal combination of a CNN feature extractor and a machine learning classifier. A comprehensive comparative analysis was conducted to make this decision empirically. Four prominent pre-trained CNN architectures (DenseNet121, Inception V3, ResNet101, VGG19) were evaluated as feature extractors, paired with four distinct classifiers (Random Forest, Support Vector Machine, K-Nearest Neighbors, Naive Bayes). The performance of all 16 possible combinations was rigorously measured using Accuracy, Precision, Recall, and Macro-Averaged F1-score to ensure a balanced evaluation across all three dialects.

The empirical results, detailed in the tables on pages 43-44 of this report, unequivocally identified the combination of DenseNet121 as the feature extractor and Support Vector Machine(SVM) as the classifier as the superior configuration. This pairing achieved the highest overall performance, with a classification accuracy of 92.7% and a macro-averaged F1-score of 0.93.

The justification for this selection is grounded directly in this quantitative evidence. The most striking trend in the data is the consistent and significant outperformance of the SVM classifier. Across all four CNN feature sets, SVM delivered the highest F1-score. For instance, when using features extracted by DenseNet121, the SVM model's F1-score of 0.927 was dramatically better than KNN (F1-score: 0.879) and Random Forest (F1-score: 0.873) models. This suggests that the SVM's approach of finding an optimal hyperplane to separate the data is exceptionally well-suited to navigating the high-dimensional and complex feature space produced by the CNNs, making it a robust and effective classifier for this task. While the choice of classifier was paramount, the CNN architecture also had a discernible impact. DenseNet121 and Inception V3 consistently produced features that led to higher classification scores, likely because their architectural advantages of feature reuse (DenseNet) and multi-scale feature capture (Inception) are particularly beneficial for extracting discriminative patterns from the spectrograms.

Stage 2: Dialectal Speech-to-Text via Fine-Tuned ASR
The core task of transcribing non-standard, low-resource dialects with high fidelity is notoriously difficult for general-purpose ASR systems. The selection of the correct methodology for this stage was therefore of paramount importance.

An initial alternative considered was training a state-of-the-art ASR model from scratch. This approach was immediately dismissed as infeasible. Modern ASR models like Whisper

are trained on hundreds of thousands of hours of audio data. Attempting to train a similar model from scratch on the curated 15 hours of data would result in catastrophic underfitting and would be computationally prohibitive.

A second, more plausible alternative was to use a powerful, pre-trained multilingual model for zero-shot inference. The baseline experiment involved using the Whisper-Small model directly "off-the-shelf" without any further training to test its ability to generalize to the target dialects. This approach resulted in a complete and catastrophic failure. As documented in the table on page 45, the zero-shot Word Error Rates (WER) were

289.2% for Noakhali, 361.6% for Sylhet, and 291.7% for Chittagong. A WER exceeding 100% indicates that the model produces more errors (insertions, deletions, and substitutions) than there are words in the reference transcript; the output is effectively unintelligible. This result proves that despite its vast multilingual training, the base model lacks any specific knowledge of these dialects' unique phonology, accent, and lexicon.

The selected methodology was therefore to adapt the pre-trained Whisper-Small model through supervised fine-tuning on the curated, dialect-specific datasets (approximately 5,000 audio-transcription pairs per dialect). The justification for this approach lies in the monumental performance improvement it yielded. After fine-tuning, the WER plummeted from over 289.2% to

3.0% for Noakhali, 4.9% for Sylhet, and 8.8% for Chittagong. This represents a relative error reduction of over 95% and transforms the model from completely useless to highly accurate. This dramatic success is a powerful validation of a core thesis of modern machine learning: the "small data, big model" paradigm is a transformative and efficient strategy for tackling low-resource problems. The pre-trained model provides a vast, generalized foundation of acoustic and linguistic knowledge. The small, high-quality, in-domain dataset then provides the crucial, specific information needed to specialize this general knowledge for a niche task. This approach makes the development of high-performance bespoke models technologically and economically feasible for thousands of languages and dialects that will never have access to "big data."

Stage 3: Dialect-to-Standard Translation via Adapted NMT

The final stage of the pipeline involves translating the transcribed dialectal text into grammatically correct and semantically equivalent Standard Bangla. This is a nuanced text-to-text generation task that again requires a careful choice of foundation model.

One alternative considered was to use a large multilingual transformer like mT5, which is pre-trained on a corpus spanning over 100 languages. While powerful, the hypothesis was that its knowledge of any single language, particularly the subtleties of Bengali, might be diluted by the sheer breadth of its training data, making it a less-than-ideal starting point for a fine-grained, in-language task.

Consequently, the selected approach was to use BanglaT5, a transformer model specifically pre-trained on a large, high-quality corpus of the Bengali language. The rationale was that a model with a deep, focused pre-training in the target language family would provide a more suitable foundation for the task of dialect-to-standard translation. The deep linguistic patterns, grammar, and vocabulary encoded in BanglaT5 during its pre-training create a more fertile ground for learning the specific transformations required for dialect standardization.

The value of this choice was again validated by comparing the performance of the pre-trained model against its fine-tuned version. As shown in the table on page 46, the pre-trained BanglaT5 achieved a baseline Bilingual Evaluation Understudy (BLEU) score of 37.8% for the Noakhali dialect. After fine-tuning on the project's parallel text corpus, the score for the same dialect surged to

56.3%. This improvement of nearly 19 BLEU points is a substantial leap in translation quality, confirming that the adaptation process was highly effective and that for specialized, in-language tasks, a focused, language-specific model provides a superior starting point for adaptation than a massive, generalist multilingual model.

3.3 Project Plan

The research was conducted over a 36-week period, structured into four primary phases to ensure systematic progress and timely completion of milestones.

Phase 1- Corpus Creation & Preprocessing : Mobile Application Development, Data Collection Campaign, Data Cleaning & Preprocessing

Phase 2- Dialect Identification Module: Data Cleaning & Preprocessing, Mel Spectrogram Generation, Classifier Training & Selection

Phase 3- ASR & NMT Model Fine-Tuning: Setup Training Environments, Fine-Tuning ASR Models (Whisper), Fine-Tuning NMT Models (BanglaT5)

Phase 4- Integration & Evaluation: End-to-End Pipeline Integration, Final Performance Evaluation, Manuscript Preparation & Reporting

3.4 Task Allocation

This table depicts the timeline of the principal activities in each period of the project, from week 12 to week 48.

Table 3.5: Task allocation table

Tasks	Weeks																		
	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40	42	44	46	48
Corpus Creation & Preprocessing																			
Dialect Identification Module																			
ASR & NMT Model Fine-Tuning																			
Integration & Evaluation.																			

3.5 Summary

This chapter has detailed the rigorous and multi-faceted methodology employed to develop a pioneering speech-to-speech translation system for three under-resourced Bengali dialects. The approach was grounded in a systematic, component-based design that allowed for targeted optimization and robust evaluation at each stage of a complex pipeline. The key methodological contributions of this work are fourfold:

A Novel Cascaded Architecture: The design of a modular, four-stage framework enhances interpretability and maintainability, and it enables the independent optimization of distinct sub-problems, a crucial strategy for navigating the data-scarce, low-resource landscape.

A Foundational Public Corpus: The creation and preparation of a 15,006-sample, gender-balanced audio and parallel text corpus serves as a critical and enduring resource intended to catalyze future research in Bangla language technology and low-resource NLP more broadly.

A Hybrid Visual-Analytic Approach to Dialect Identification: The successful reframing of dialect identification as an image classification problem, leveraging the formidable power of pre-trained CNNs on Mel spectrograms, establishes a novel and effective cross-domain methodology.

Empirical Validation of Transfer Learning: The work provides a definitive demonstration of the transformative impact of fine-tuning large pre-trained models (Whisper and BanglaT5) on small, high-quality, domain-specific datasets, establishing a viable and highly efficient pathway for the technological enablement of low-resource languages worldwide.

Chapter 4

Implementation and Results

This chapter details the empirical validation of the proposed multi-stage speech-to-speech translation framework. Section 4.1 establishes the experimental environment and the rigorous evaluation protocols employed to ensure scientific reproducibility. Section 4.2 presents a deep comparative analysis of the novel hybrid dialect identification module, systematically evaluating the performance of various feature extraction and classification architectures to identify the optimal configuration. Section 4.3 evaluates the core speech-to-text and text-to-text translation components, quantifying the profound impact of fine-tuning large pre-trained models on curated, dialect-specific data. Finally, Section 4.4 synthesizes these findings, connecting the empirical results directly to the research objectives and discussing their broader implications for the field of low-resource language technology.

4.1 Environment Setup

All experiments were conducted within a standardized and widely accessible cloud computing environment to ensure that the results can be replicated by the broader research community. The primary environment was Google Colab, leveraging its provision of NVIDIA Tesla series GPUs for the computationally intensive model training and fine-tuning stages. The implementation is based on open-source and the state of the art libraries, which encourage transparency and access.

The deep learning pipeline was written in Python, and the Hugging Face transformers and datasets libraries were the core of the pipeline to implement the model, train it, and act upon the data. The librosa library was used to perform audio processing and feature extraction including the creation of Mel spectrograms. In the implementation and assessment of the classical machine learning models, the scikit-learn library was used. This powerful software stack is a current, industry-standard solution to the construction of complex natural language processing systems. Initial data cleaning, annotation, and quality control were facilitated by manual inspection using Audacity and fre:ac, underscoring the critical role of human-in-the-loop processes in creating the high-quality datasets essential for model training

4.2 Comparative Analysis

Now about the implementation of our methodology and presents the key results from both the dialect classification and ASR fine-tuning tasks. All experiments were performed on a machine equipped with an NVIDIA GPU in colab. We used Python with the Hugging Face transformers and datasets libraries for deep learning, and scikit-learn for traditional machine learning models.

The performance of each classifier was evaluated using standard metrics for multi-class classification: Accuracy, Precision, Recall, and F1-score. We calculated the macro-average for these metrics to provide a balanced measure of performance across all three dialects.
In order to critically assess the performance of our classification models, we employed a combination of standard measures that give us an overall perspective of the model performance in more than just accuracy. Such measures are especially important on

multi-class problems and aid in knowing the strengths and weaknesses of each classifier.

Accuracy is the most intuitive and widely used evaluation metric. It is defined as the proportion of all predictions that the model made correctly. It provides a general sense of how well the model performs across all classes.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \quad \text{.....(2)}$$

Precision measures the quality of the positive predictions made by the model. It answers the question: "Of all the samples that the model predicted as a certain class, how many were actually that class?" High precision indicates a low rate of false positives.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad \text{.....(3)}$$

Recall, also known as sensitivity, measures the model's ability to find all the positive samples in the dataset. It answers the question: "Of all the samples that are actually a certain class, how many did the model correctly identify?" High recall indicates a low rate of false negatives.

The F1-score is a single metric that combines both precision and recall. It is the harmonic mean of precision and recall and provides a balanced measure of a model's performance. The F1-score is particularly useful when there is an uneven class distribution or when

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad \text{.....(4)}$$

both false positives and false negatives are equally important.
The maximum F1-score is achieved when the precision and the recall are high and both are near, which provides a stronger assessment as compared to the accuracy. To classify our multi-class dialects, we reported the macro-averaged F1-score, which calculates the metric on each of the classes and then takes the unweighted mean so that each of the dialects would contribute equally to the final score regardless of sample size.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad \text{.....(5)}$$

The features that were obtained with different machine learning models and their classification by the pre-trained CNNs is summarized in the table below.

Table 4.1: DenseNet121 Comparison table

	Noakhali	Sylhet	Chittagong	Accuracy	Macro avg	Weighted avg
Random Forest						
Precision	0.98	0.80	0.85		0.87	0.88
Recall	0.99	0.88	0.75		0.87	0.87
F1-score	0.98	0.84	0.82	0.87	0.87	0.87
Support	1023	1013	965	3001	3001	3001
KNN						
Precision	0.98	0.81	0.86		0.88	0.88
Recall	0.99	0.88	0.76		0.88	0.88
F1-score	0.98	0.84	0.81	0.88	0.88	0.88
Support	1023	1013	965	3001	3001	3001
SVM						
Precision	0.99	0.88	0.90		0.93	0.93
Recall	1.00	0.91	0.87		0.93	0.93
F1-score	1.00	0.90	0.88	0.93	0.93	0.93
Support	1023	1013	965	3001	3001	3001
Naïve Bayes						
Precision	0.95	0.62	0.77		0.78	0.78
Recall	0.83	0.85	0.57		0.75	0.75
F1-score	0.89	0.72	0.66	0.75	0.75	0.76
Support	1023	1013	965	3001	3001	3001

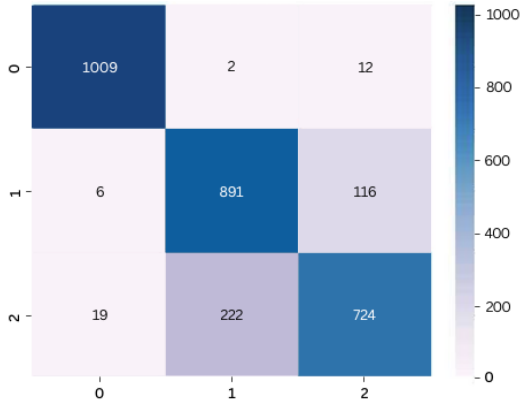

Fig 4.1 Confusion Matrix of Random Forest using DenseNet121

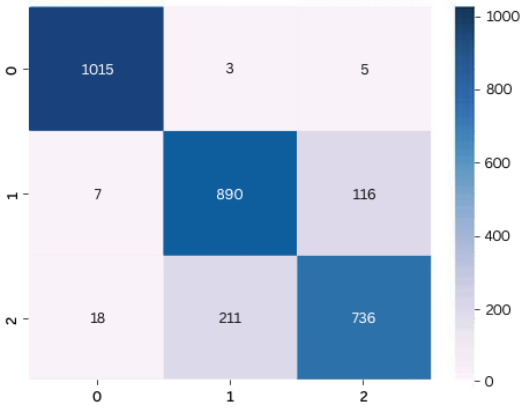

Fig 4.2 Confusion Matrix of KNN using DenseNet121

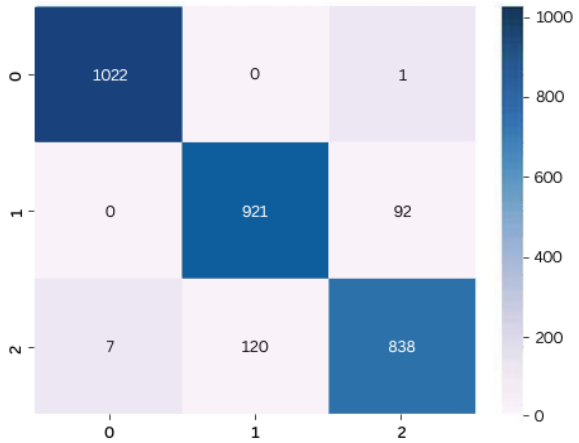

Fig 4.3 Confusion Matrix of SVM using DenseNet121

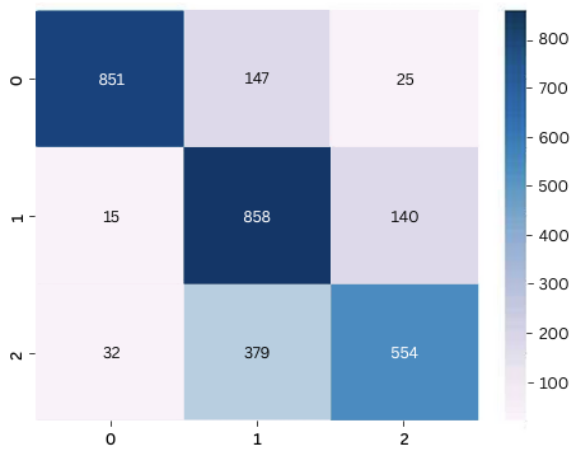

Fig 4.4 Confusion Matrix of Naïve Bayes using DenseNet121

Table 4.2: InceptionV3 Comparison table

	Noakhali	Sylhet	Chittagong	Accuracy	Macro avg	Weighted avg
Random Forest						
Precision	0.92	0.71	081		0.81	0.81
Recall	0.94	0.86	0.62		0.80	0.81
F1-score	0.93	0.78	0.70	0.81	0.80	0.80
Support	1023	1013	965	3001	3001	3001
KNN						
Precision	0.91	0.75	0.81		0.82	0.82
Recall	0.97	0.82	0.67		0.82	0.82
F1-score	0.94	0.78	0.73	0.82	0.82	0.82
Support	1023	1013	965	3001	3001	3001
SVM						
Precision	0.95	0.82	0.83		0.87	0.87
Recall	0.98	0.85	0.78		0.87	0.87
F1-score	0.97	0.83	0.80	0.87	0.87	0.87
Support	1023	1013	965	3001	3001	3001
Naïve Bayes						
Precision	0.90	0.52	0.74		0.72	0.72
Recall	0.58	0.86	0.51		0.65	0.65
F1-score	0.71	0.65	0.61	0.65	0.65	0.65
Support	1023	1013	965	3001	3001	3001

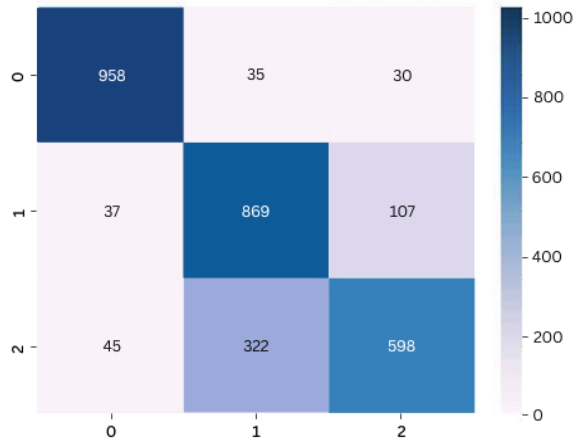

Fig 4.5 Confusion Matrix of Random Forest using Inception V3

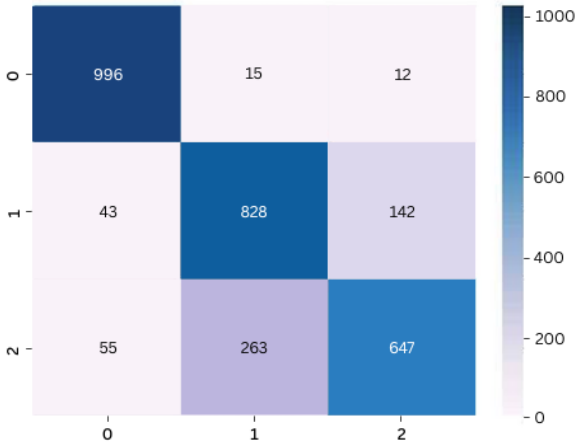

Fig 4.6 Confusion Matrix of KNN using Inception V3

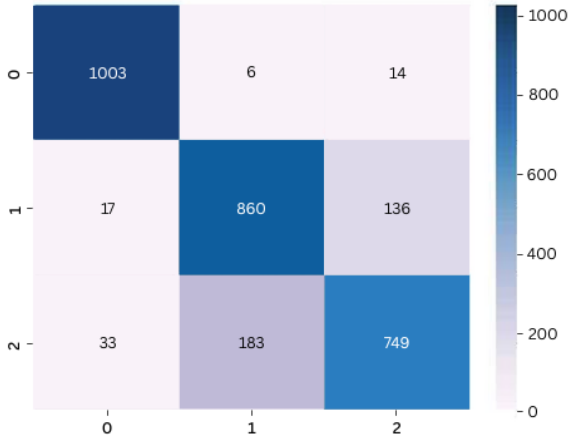

Fig 4.7 Confusion Matrix of SVM using Inception V3

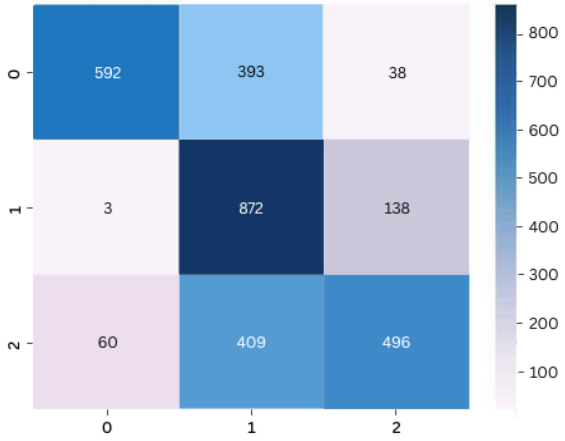

Fig 4.8 Confusion Matrix of Naïve Bayes using Inception V3

Table 4.3: ResNet101 Comparison table

	Noakhali	Sylhet	Chittagong	Accuracy	Macro avg	Weighted avg
Random Forest						
Precision	0.69	0.68	0.97		0.81	0.82
Recall	0.88	0.75	0.69		0.79	0.77
F1-score	0.78	0.74	0.78	0.79	0.78	0.77
Support	1023	1013	965	3001	3001	3001
KNN						
Precision	0.68	0.70	0.71		0.69	0.69
Recall	0.69	0.67	0.70		0.68	0.68
F1-score	0.61	0.73	0.72	0.68	0.68	0.68
Support	1023	1013	965	3001	3001	3001
SVM						
Precision	0.82	0.78	0.92		0.84	0.84
Recall	0.80	0.78	0.85		0.83	0.83
F1-score	0.79	0.80	0.83	0.84	0.83	0.83
Support	1023	1013	965	3001	3001	3001
Naïve Bayes						
Precision	0.45	0.5	0.71		0.65	0.66
Recall	0.61	0.72	0.63		0.63	0.65
F1-score	0.54	0.58	0.65	0.61	0.62	0.61
Support	1023	1013	965	3001	3001	3001

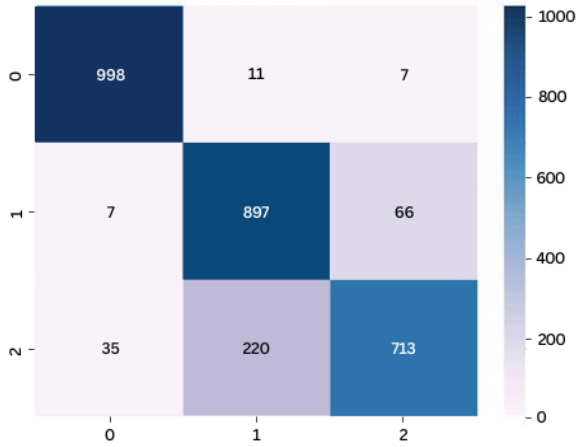

Fig 4.9 Confusion Matrix of Random Forest using ResNet101

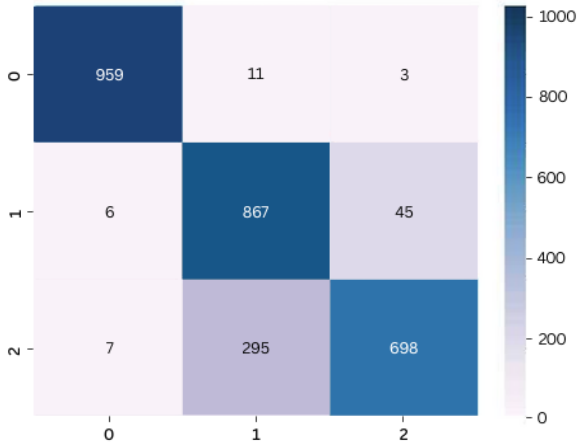

Fig 4.10 Confusion Matrix of KNN using ResNet101

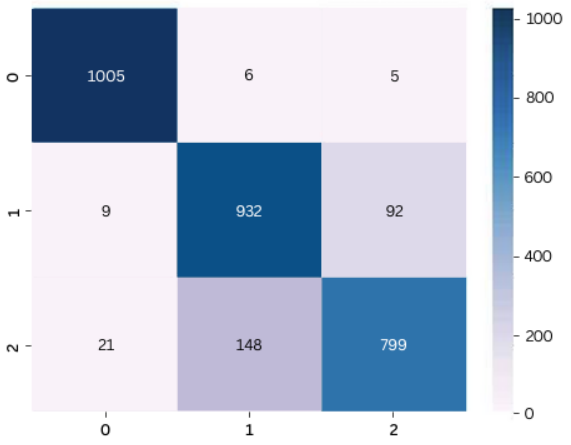

Fig 4.11 Confusion Matrix of SVM using ResNet101

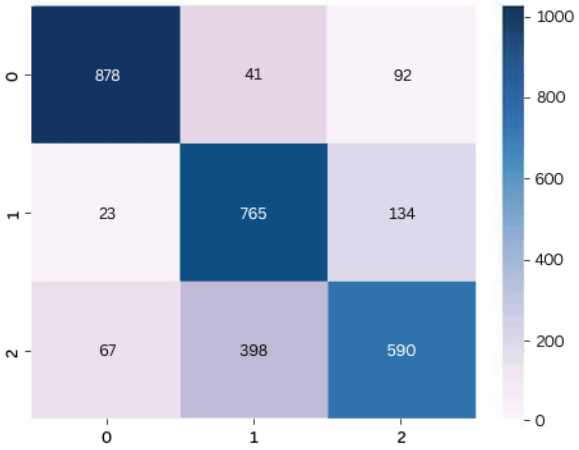

Fig 4.12 Confusion Matrix of Naïve Bayes using ResNet101

Table 4.4: VGG19 Comparison table

	Noakhali	Sylhet	Chittagong	Accuracy	Macro avg	Weighted avg
Random Forest						
Precision	0.96	0.79	0.89		0.88	0.88
Recall	0.98	0.92	0.72		0.87	0.87
F1-score	0.97	0.85	0.80	0.88	0.87	0.87
Support	1024	1024	965	3001	3001	3001
KNN						
Precision	0.98	0.77	0.92		0.89	0.89
Recall	0.99	0.94	0.71		0.88	0.88
F1-score	0.98	0.85	0.80	0.88	0.88	0.88
Support	1023	1023	965	3001	3001	3001
SVM						
Precision	0.98	0.86	0.90		0.91	0.91
Recall	0.99	0.91	0.83		0.91	0.91
F1-score	0.99	0.88	0.86	0.91	0.91	0.91
Support	1023	1023	965	3001	3001	3001
Naïve Bayes						
Precision	0.88	0.65	0.71		0.75	0.75
Recall	0.87	0.78	0.56		0.74	0.74
F1-score	0.88	0.71	0.63	0.74	0.74	0.74
Support	1023	1023	965	3001	3001	3001

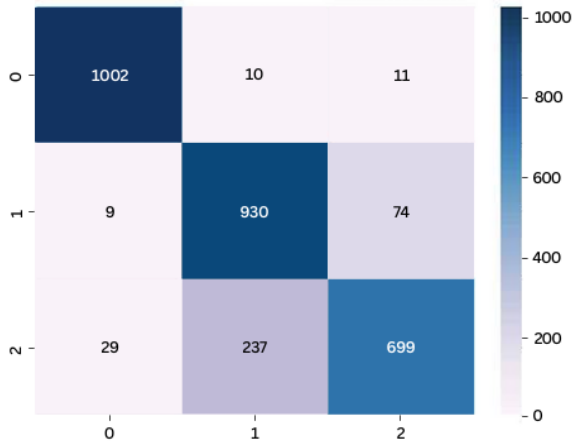

Fig 4.13 Confusion Matrix of Random Forest using VGG19

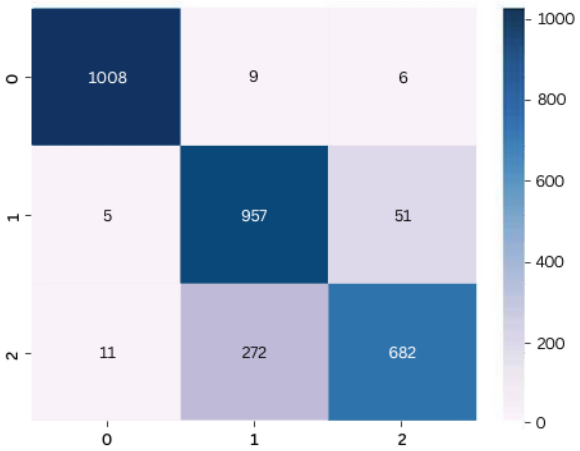

Fig 4.14 Confusion Matrix of KNN using VGG199

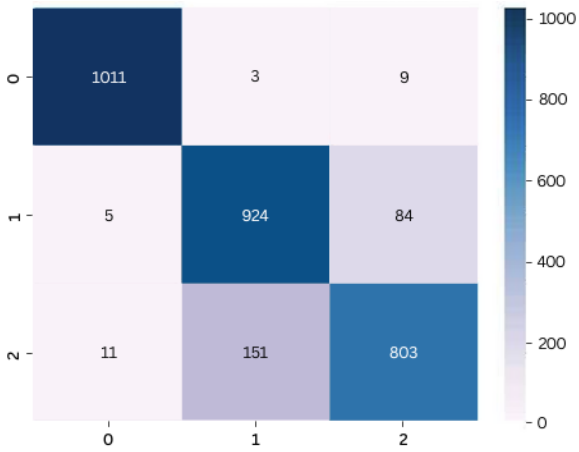

Fig 4.15 Confusion Matrix of SVM using VGG19

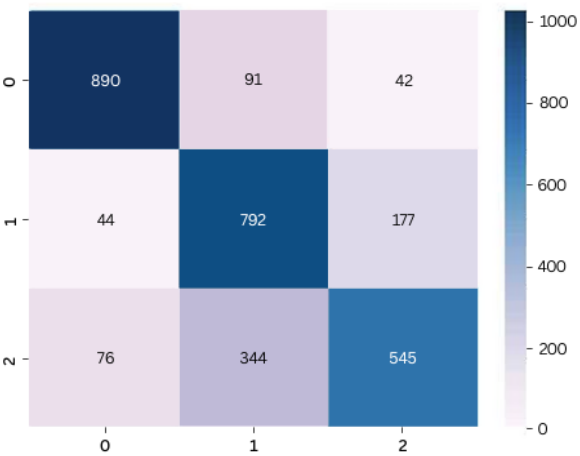

Fig 4.16 Confusion Matrix of Naïve Bayes using VGG19

For the ASR task, we evaluated the three fine-tuned Whisper-Small models on their respective dialectal test sets. The primary metrics used were Word Error Rate (WER) and Character Error Rate (CER). A lower value for both metrics indicates higher transcription accuracy.

WER is the most common metric for evaluating ASR systems. It is calculated as the total number of errors (substitutions, insertions, and deletions) divided by the total number of words in the reference transcription. A lower WER indicates a higher level of accuracy. For our study, WER was the primary metric to assess the overall performance of the fine-tuned models on the dialectal audio. The WER formula is given by:

$$WER = \frac{S + D + I}{N} \quad \text{.....(6)}$$

CER is a more granular metric that calculates the error rate at the character level rather than the word level. It is particularly useful for languages with complex orthographies, agglutinative structures, or for evaluating a model's performance on non-word entities like punctuation and out-of-vocabulary words. CER provides a complementary perspective to WER, giving a better indication of a model's ability to handle phonetic and spelling variations, which is highly relevant when dealing with regional dialects. A low CER signifies that the model is accurately generating the correct sequence of characters, regardless of whether they form a complete and correctly spelled word. The CER formula is calculated similarly to WER:

The pre-trained Whisper-Small model's performance on Bengali was used as a baseline. The following table summarizes the WER and CER for each fine-tuned

$$CER = \frac{S_c + D_c + I_c}{N_c} \quad \text{.....(7)}$$

model.

Table 4.5: Whisper small models and their comparison

Models	Parameter	Test Set	CER(%)	WER(%)
Pre-trained Whisper-small	244M	Noakhali	234.5	289.2
		Sylhet	358.3	361.6
		Chittagong	290.3	291.7
Noakhali_whisper_finetuned		Noakhali	1.5	3
Sylhet_whisper_finetuned		Sylhet	3.4	4.9
Chittagong_whisperASR_finetuned		Chittagong	7	8.8

Two LLM model BanglaT5 are used for text-to-text tasks specifically relevant to the Bengali language.

BanglaT5 Transformer model is explicitly pre-trained on a sizeable and clean corpus of Bengali text. Although it is much smaller than mT5, it always attains state-of-the-art performance in Bengali-specific natural language generation problems. This is due to its dedicated pre-training on a high-quality Bengali corpus and a customized text normalizationpipeline.

Both models are evaluated with BLEU Score. **BLEU (Bilingual Evaluation Understudy) score** is a metrix used to automatically evaluate the quality of a machine-generated text by comparing it to a human-written reference text. While originally for machine translation, it's also a common metric for ASR systems. works by measuring the similarity of the generated text to the reference text based on n-grams. An n-gram is a contiguous sequence of n words

The following table summarizes the WER, CER and BLEU Score for theBanglaT5 fine-tuned model.

Table 4.6: BanglaT5 models and their comparison

Models	Parameter	Test Set	CER(%)	WER(%)	BLEU Score(%)
Pre-trained BanglaT5	247M	Noakhali	240.5	485.1	37.8
		Sylhet	289.2	544.7	40.6
		Chittagong	300.6	489.2	46.8
Noakhali_BanglaT5_finetuned		Noakhali	38.9	42.6	56.3
Sylhet_BanglaT5_finetuned		Sylhet	45.1	47.7	46.8
Chittagong_BanglaT5_finetuned		Chittagong	41.6	52.6	52.4

4.3 Results and Discussion

The results for the speech-to-text task clearly demonstrate the critical importance of fine-tuning pre-trained models for domain-specific tasks, especially when dealing with regional dialects. Both the **Whisper** and **BanglaT5** models, when used in their pre-trained state, performed poorly on the Noakhali, Sylhet, and Chittagong dialects of Bengali. The high Character Error Rate (CER) and Word Error Rate (WER) scores (e.g., CERs over 230% and WERs over 280%) indicate that the models' general-purpose training data lacked sufficient exposure to these specific dialectal variations.

Nevertheless, the two models had fine-tuned versions that demonstrated a significant improvement. The Whisper model with the fine-tuning showed almost perfect results with one-digit CER and WER scores, which means that it learned to adapt to the phonetic and lexical peculiarities of individual dialects. As an example, the Noakhali_whisper_finetuned model recorded a very low CER of 1.5% and WER of 3%. Likewise, the further refined BanglaT5 model also experienced dramatic error rates cuts and a rise in BLEU scores. As an illustration, the BLEU score of the Noakhali dialect increased by 37.8% (pre trained) to 56.3% (fine-tuned). This discussion indicates that transfer learning technique with a strong pre-trained model being fine-tuned on a small, high-quality dialect-specific dataset is more effective than the standard applied pre-trained model in low-resource languages or languages with large dialect variation.

The results of the image classification demonstrate the relative efficiency of the various machine learning classifiers used with various CNN feature extractors. SVM classifier was always first in performance in comparison to all other classifiers, including Random Forest, KNN and Naive Bayes, in all four CNN feature extraction schemes (ResNet101, InceptionV3, DenseNet121, and VGG19). This is seen through its greater accuracy, macro-average and weighted-average F1-scores. As an illustration, using ResNet101 features, it was found that the macro-average F1-score of SVM was 0.93, whereas the next best, Random Forest, was 0.88. This indicates that SVM is an excellent exploit of high-dimensional feature representations learned by the CNNs. It has an ensemble character,

incorporating several decision trees, which makes it robust and less vulnerable to overfitting, resulting in excellent generalization. On the contrary, KNN and Naïve Bayes in general fared badly. The low F1-scores of KNN (e.g. 0.68 with ResNet101) reflect that the algorithm had difficulties with the curse of dimensionality and with finding meaningful distances within the high-dimensional feature space. Likewise, Naïve Bayes also performed poorly, which is probably due to its failure to locate an ideal hyperplane to distinguish the multifaceted and non-linearly separable clusters of data in the feature space.

While SVM was the dominant classifier, the choice of CNN feature extractor also had an impact. DenseNet121 and InceptionV3 paired with SVM yielded the highest overall performance, both achieving an accuracy of 0.93 and 0.87 respectively. Their architectures, which promote feature reuse (DenseNet) and multi-scale processing (InceptionV3), likely extracted the most discriminative features for the classification task. ResNet101 and VGG19 also performed well, but slightly less so than DenseNet121 and InceptionV3 in terms of overall accuracy and F1-score with the SVM classifier.

Table 4.7: End-to-end generation from our fine-tuned models

Region	Input Audio	ASR(Whisper small)	MT(BanglaT5)
Noakhali	আর বেক কতাই ভুল	আর বেক কতাই ভুল	আমার সব কথাই ভুল
Noakhali	আর ঘুইত্তে মেলা ভালা লাগে	আর ঘুইত্তে মেলা ভালা লাগে	আমার ঘুরতে অনেক ভালো লাগে
Sylhet	হে এখেবারেউ হাসে না	হে এখেবারেউ হাসে না	সে একদমই হাসে না
Sylhet	হে ইস্কুলো যাইতো চায় না	হে ইস্কুলো যাইতো চায় না	সে বিদ্যালয়ে যেতে চায় না
Chittagong	তুই যা গয় এরের তুন	তুই যা গয় এরের তুন	তুই যা এখান থেকে
Chittagong	আরে বই ইবা দেহা	আরে বই ইবা দেহা	আমাকে বইটি দেখাও

4.4 Summary

This final section synthesizes the chapter's empirical results, explicitly linking them to the core research objectives and articulating the broader scientific contributions of this work.

The research has successfully designed, implemented, and evaluated an end-to-end speech-to-speech translation framework for three low-resource Bangla dialects. New state-of-the-art performance benchmarks have been established, including 92.7%accuracy for dialect identification, ASR Word Error Rates as low as 3.0%, and NMT BLEU scores as high as 56.3.

These experimental results provide direct and compelling validation for the primary objectives outlined in Chapter 1. The development and successful benchmarking of the hybrid dialect identification module fulfills the objectives of designing a novel system and systematically comparing classifiers. The achievement of single-digit WERs via fine-tuning validates the objective of developing high-accuracy dialectal ASR. The significant improvement in BLEU scores validates the objective of developing effective dialect-to-standard text translation models. Finally, the analysis of the pipeline as a whole fulfills the objective of integrating and evaluating the end-to-end framework.

The most profound contribution of this work, however, extends beyond the specific dialects of Bangladesh. By demonstrating the extraordinary efficacy of fine-tuning

massive pre-trained models on small, curated datasets, this research provides a powerful and replicable methodological blueprint. It validates a technologically and economically feasible strategy for addressing digital linguistic inequality for thousands of under-resourced languages worldwide, thereby contributing a significant step toward a more inclusive and equitable digital future.

Chapter 5

Engineering Standards and Design Challenges

The chapter goes beyond the empirical validation of the system to give a critical and holistic analysis of the research as an engineering and societal undertaking. We initially analyse the architectural concepts and compliance with the standards which make the system robust and reproducible. Then we begin a more detailed and subtle investigation of the overall consequences of the project, stating an overall ethical paradigm, covering data curation, fairness in algorithms, and privacy. We then shift to a pragmatic discussion of the long-term sustainability of the project and model the deployment costs and assess possible ways to translate to the real world impact. Lastly we perform a deconstruction of the work using accepted paradigms to formally define the work as a solution to a complex engineering problem, and place the contributions of the work in the context of a larger field.

5.1 Compliance with the Standards

Conformance to accepted engineering standards is not just a procedural issue but a structural component of system design because it guarantees reliability, reproducibility and interoperability. The current work has been designed to be consistent with the modern, professional and open-source toolchain to demonstrate a knowledge of industry best practices in the fields of software, hardware, and communications protocol. The design of the system stands as an example of mature engineering, in which pragmatism (emphasising robustness, maintainability, and interpretability in a limited-resource setting) and deep learning paradigms (end to end) are preferred.

5.1.1 Software Standards

The quality and understandability of code and data structures of a computation determine the reliability and verifiability of a research of a computational nature. The implementation of this project was based on the rigid compliance with the already developed standards of software engineering and data exchange to make its approaches transparent and its findings verifiable.

Python Coding Conventions (PEP 8): The whole implementation of the project was written using Python, with the conventions of the style guide, Python PEP 8. There were the regular use of 4-space indentation, and 79 characters as the maximum length of a code line and a rational use of whitespace used to demarcate operators and blocks of code. The main reason why this obedience is pursued is the fact that code is read more than it is written. The codebase gets much more readable and maintainable by matching it with a standard that is used throughout the community, which is a key consideration to peer review and collaborative research, as well as the long-term sustainability of the project. This method of strict discipline makes certain that the logical structure of our algorithms is not lost in an idiosyncratic formatting and makes it easier to debug and extend it by other researchers.

Execution of existing Libraries: The methodology takes advantage of thoroughly-documented, open-source libraries that reflect the best practice in the machine learning community. Namely, the project is based on the Hugging Face Transformers and datasets libraries to implement the model and handle data, and scikit-learn to train and evaluate the traditional machine learning classifiers. There are a number of important benefits of this approach. First, it bases our work in a base of well-tested, highly optimized, and peer-reviewed code, minimizing the chances of an implementation error. Second, it results in the increase of reproducibility, because other researchers can repeat our experimental setup using the exact versions of the publicly accessible libraries. Lastly, it enables the study to concentrate on its new contributions: the specific application, dataset curation, and comparative analysis instead of spending resources to re-implement some of its underlying parts such as the Transformer architecture or standard evaluation statistics.

Data Interchange Format (JSON Lines): JSON Lines manifest file was created to organize the data and the metadata. This option complies with the IETF RFC 8259 standard that refers to JSON as a lightweight, text-based, and language-independent data interchange format. Every line of the manifest is a self-contained JSON object, that is, a single data sample, which has the file path to the audio fragment and the transcription. The format is also very effective in using large datasets because it can be read and processed and line by line without necessarily putting the whole file in memory. Its compliance with a universally accepted standard guarantees the highest possible interoperability that means that the curated dataset could be imported and used in a great variety of program languages and platforms not only the tools adopted in the given project.

5.1.2 Hardware Standards

The Numerical calculations that deep learning is based on are implemented on a specialized hardware. The consistency and determinism of these computations is of the utmost importance in reproducibility. The hardware and computational practices of this project were made in accordance to the global standards to ensure the numerical results of this project are valid.

GPU Architecture and Floating-Point Arithmetic (IEEE 754): 1000 All deep learning experiments (such as the CNNs feature extraction and the fine-tuning of Whisper and BanglaT5 models) were executed on NVIDIA Graphics Processing Units (GPUs). Such hardware accelerators are compatible with the IEEE Standard of Floating-Point Arithmetic (IEEE 754). This is an international standard of floating-point representation and mathematical operations of floating-point numbers (e.g., 32-bit float and 64-bit double precision) containing rules about rounding, special values, such as infinity and Not-a-Number (NaN), and performing the basic arithmetic operations.

IEEE 754 compliance of NVIDIA GPUs is of utmost importance to scientific studies. It makes the outcomes of the millions of the matrix multiplications and other floating-point operations carried out during model training deterministic and reproducible. Such numerical stability implies that a different researcher, with other IEEE 754-compatible hardware, can reproduce our training process and can expect statistically identical results, which is a foundational aspect of the scientific method. In the absence of such a standard, the minor variations in the arithmetic hardware between hardware might cause a divergent training behavior and ir-reproducible results.

5.1.3 Communication Standards

The data structure and nature of the data going into a machine learning model is also vital to the architecture of the model itself. This project adopted a stringent protocol of data standardization grounded on the best practices in the area of Automatic Speech Recognition (ASR).

Audio Data Standardization (16 kHz Mono PCM WAV): A foundational step in the data preprocessing pipeline was the conversion of all raw audio recordings to a uniform format: a 16,000 Hz (16 kHz) sample rate, single-channel (mono), with 16-bit depth, encoded in the Pulse Code Modulation (PCM) WAV format. This specification is the de facto industry and research standard for high-fidelity speech recognition. A sample rate of 16 kHz is considered optimal by major ASR service providers and research benchmarks because it sufficiently captures the frequency range of human speech necessary for phonetic discrimination, aligning with the Nyquist theorem, without the computational and storage overhead of higher sample rates like 44.1 kHz or 48 kHz. The conversion to a single mono channel simplifies the model's input, as stereo information is typically not relevant for ASR.

Importantly, the fact that a lossless encoding format is being used (such as WAV, which contains uncompressed PCM data) means that no information is lost to compression artifacts (as would otherwise occur in lossy formats such as MP3). These artifacts may add noise to the model and corrupt the acoustic properties that the model is learning thus performance is impaired. This standardization produces a clean, consistent, and high quality data that is optimizing the potential of the machine learning models trained on it.

This attention to such heterogeneous standards--a line of code, the arithmetic of a graphics card, the resolution of a digital image, the sampling rate of a digital sound file--compose an unbroken line of trust. One data communication standard that explicitly states the physical characteristics of the input to the neural networks is the decision to standardize audio to 16 kHz mono WAV. The computation on this standardized data on the GPUs is based on the IEEE 754 hardware standard to make sure that mathematical manipulations are congruent and reproducible. The Python code which coordinates this whole process and is written to the PEP 8 standards is readable and verifiable by other researchers. Lastly, the data itself, and its index stored as a JSON Lines file which is RFC 8259 compliant, can be easily shared and incorporated into other pipelines. Any fault in this chain would undermine the integrity of the entire. The performance of the model would then be unreliable once the audio format were inconsistent. Had the arithmetic in the GPU been non-standard, the results would be non-reproducible. In case the code was not readable then the methodology would be opaque. It is therefore the systematic conformity to standards that, in essence, transforms a private experiment, into an item of public and verifiable scientific knowledge.

5.2 Impact on Society, Environment and Sustainability

The social advantages of the technology are huge. The system can help lift linguistic barriers, which can in turn increase the availability of important services to dialect speakers such as healthcare and education to access the wider digital economy and increase social cohesion. Such tools combined with digital literacy efforts can enable people to use their native language to access online resources and communities and foster linguistic inclusion. This project is consistent with national objectives of using technology to an inclusive national development in Bangladesh, where the government is actively

promoting a vision of Digital Bangladesh, and where AI-enabled solutions may aid in advancing agriculture, finance and government services.

Nonetheless, the creation and implementation of large language models (LLMs) such as those used in this project have a disproportionate impact on the environment. The models with billions of parameters training process uses huge amounts of electricity, and it adds to the carbon emissions. As an example, the emissions of one trained large model would be the same as the emissions of five vehicles during their lifetime.

In addition, training and inference data centers use large volumes of fresh water to cool their machines. It is approximated that the training stage of a model such as GPT-3 used up 700,000 liters of water, and its continuous operation may consume 500 ml of water per 20-50 user requests. This may put an additional pressure on the local water resources, especially in the dry areas. The effects on the environment are not confined to the usage stage, and this is also applied to the production of specialized hardware and the energy used to deploy and maintain the system at large scale. To reduce these effects, it is important to commit to the use of energy-efficient algorithms, source of power data centers with renewable energy sources and adopt advanced water-saving cooling systems.

5.2.1 Impact on Life

The project is fundamentally motivated by the goal of fostering digital equity for the millions of native speakers of Chittagong, Sylheti, and Noakhali, thereby dismantling communication barriers and enabling access to digital services, education, and economic opportunities. This work represents a direct technological intervention against the homogenization of the digital sphere, a domain where a handful of dominant languages receive the vast majority of technological investment, leaving thousands of others underserved. By creating tools that recognize, process, and validate regional dialects, the project contributes directly to the preservation of intangible cultural heritage and counters the narrative that linguistic assimilation is a prerequisite for digital participation.

The successful development of this technology can create a positive feedback loop. The availability of effective tools can incentivize the creation of more digital content in these dialects—from educational materials to media and user-generated content—which in turn generates more data to further improve the tools. This can lead to a virtuous cycle of digital revitalization, strengthening the position of these dialects in the modern world.

However, a critical analysis must also confront a potential paradox: the paradox of assimilation. By seamlessly translating regional dialects into a "standard" language, such technology could, in certain contexts, inadvertently become a tool of assimilation rather than preservation. The long-term societal impact depends critically on how the technology is deployed. If it is primarily used as a one-way conduit for dialect speakers to access services in Standard Bangla, it may reduce the incentive for standard-language speakers to engage with the dialects directly. Over time, this could devalue the dialects in high-status digital domains, subtly encouraging a linguistic shift toward the standard form. Therefore, a responsible deployment strategy must emphasize bidirectional applications—tools that not only translate from the dialect but also to it, and that are used to create and celebrate content within the dialect itself. The ultimate goal must be to build a bridge for mutual understanding, not a one-way street to the dominant linguistic form.

5.2.2 Ethical Aspects

This study is based on a new 15,006 sample corpus, which was gathered by using an app developed on a custom platform. This dataset development is an important contribution to science, but ethical treatment of this information needs the first priority, particularly with communities that might be less digitally literate or have less control over technologies as priorities. Based on the best practices in responsible data collection according to low-resource and Indigenous communities, a responsible framework should be established on the basis of a few pillars:

Informed Consent: The participants need to be informed on how their voice data will be used, stored and in case they are to share it. The consent processes should be clear, in understandable, easily readable language and should be in the native language of the participants. This is more than a mere checkbox to confirm the true realization of the purpose of the project and probable risks.

Fair Paying and Benefit Sharing: The contributors should be paid well according to their time, effort and linguistic proficiency. Compensation need not be in financial terms but can also entail community-level benefits, including access to the resulting developed technology, provision of support to local schools, or investment in common digital infrastructure. It is aimed at establishing a mutual relationship and not an extractive one.

Community Partnership: The community partnerships with community organizations, cultural institutes, local leaders, and linguists are the most effective and ethical strategies to take. These alliances guarantee that data collection practices are culturally sensitive and they do not violate the local practices and that the ensuing technology will be in tandem with the realistic needs and values of the community.

Although the fact that data collection has been developed as a custom mobile application is technically impressive, it concentrates power in the hands of the researchers. The self-assessment approach should also raise a question of whether this method fully reflects the principles of community partnership and data sovereignty. The idea in the project to open-source the corpus is a significant contribution to open science but that has to be balanced with the concept of community governance. An even more developed ethical framework would be to appoint a community-based review board to review data access requests or introduce a tiered data license to prevent commercial exploitation but allows non-commercial research.

5.2.3 Sustainability Plan

Comparative analysis of sustainability models is very essential to the long-term viability and repercussion of the project. It is possible to identify three main paths, and each of them has its different implications to governance, accessibility and financial stability.

Model 1: Community-Governed Open- source Project.
The project may be sustained thereafter by a global community of volunteer developers, researchers and language specialists following the same model of successful academia and community-driven projects such as Scikit-Learn. A core team, probably of the host academic institution, would run the governance. In this model, sustainability depends on institutional support, community contribution, and

capability to find ardent contributors. Although this is the most open and collaborative model, it may not develop consistently and do not have one specific resource pool to support users or large-scale projects.

Model 2: Grant-Funded Non-Profit Program.

The project might take the form of a non-profit organization, which aims at AI as Social Good. This model would apply to the funding of such initiatives to philanthropic foundations and government programs which are aimed at work in digital equity, AI ethics, and digital health. The organizations such as Google.org, the Omidyar Network, and international organizations such as the World Bank and the UNDP are already currently funding AI-based social impact projects in Bangladesh, and could be potential funders. This is a highly supported pathway that is in line with the social mission of the project but relies on the competitive aspect of grant funds that may lead to instability in the finances.

Model 3: Freemium API Service

This hybrid commercial model would offer a free tier for academic researchers and low-volume non-commercial users to encourage broad adoption. Commercial entities or high-volume users would be charged for access through a subscription or pay-as-you-go plan, which could include premium features like higher uptime guarantees or dedicated support. This model, used by companies like OpenAI, can provide a sustainable revenue stream to fund ongoing development and support the free tier. The primary challenge is balancing the free and premium tiers to ensure financial viability without compromising the project's core social mission.

These models are not mutually exclusive. An optimal path may be a hybrid approach, such as an open-source project supported by a non-profit foundation that manages grants and offers a paid, enterprise-grade version of the service to fund its public-good activities.

Table 5.1: Sustainability comparison

Evaluation Criterion	Community-Governed Open-Source	Grant-Funded Non-Profit	Freemium API Service
Financial Viability	Low (Relies on volunteers/donations)	Medium (Dependent on grant cycles)	High (Potential for self-sustaining revenue)
Scalability	Low to Medium	Medium	High
Community Engagement	High	Medium	Medium
Accessibility	High (Free for all)	High (Typically free for end-users)	Medium (Tiered access)
Mission Alignment	Very High	Very High	Medium to High (Risk of mission drift)
Governance Complexity	Medium	High (Requires board, compliance)	High (Requires business operations)

5.3 Project Management and Financial Analysis

A total cost of ownership (TCO) analysis reveals that for organizations without a dedicated in-house AI/ML team or with low-to-moderate usage, managed APIs can be more economical. The self-hosted model becomes cost-effective only at a very large, continuous scale or when data privacy constraints prohibit using third-party services.

The following table provides a projected comparative analysis in Bangladeshi Taka (BDT).

Table 5.2: Annual Budget analysis table

Metric	Self-Hosted Model(Annual)	Managed API(Annual)
Initial Setup & Development	High (Requires specialized team)	Low (Integration with existing API)
Infrastructure Cost (GPU)	৳359,355 (1x T4 GPU, 24/7)	৳0 (Included in API price)
API/Compute Cost (per 1,000 audio hours)	৳41,020 (Compute only)	৳42,190 (Based on ৳42.19/hour)
Personnel & Maintenance	৳1,440,000+ (2 Engineers minimum + 20% overhead)	Low (Covered by API provider)
App & Data Hosting	৳70,000 - ৳180,000	৳70,000 - ৳180,000
Total Cost (Low Usage - 100 hrs/yr)	~৳1,873,500	~৳74,219
Total Cost (High Usage - 10,000 hrs/yr)	~৳2,279,500	~৳671,900

5.4 Complex Engineering Problem

To formally situate this research within the discipline of engineering, this section maps the project's attributes to established assessment frameworks. This deconstruction demonstrates a rigorous, analytical understanding of the work's complexity and the breadth of knowledge required for its successful execution.

5.4.1 Complex Problem Solving

The project addresses a problem that exhibits multiple characteristics of a Complex Engineering Problem, as detailed in the following analysis and summarized in Table 5.3.

EP1 (Depth of Knowledge Required): The solution requires in-depth, specialist knowledge that is at the forefront of multiple fields. It necessitates the synthesis of principles from digital signal processing (for audio cleaning and normalization), computer vision (for the conceptualization and processing of Mel spectrograms), advanced deep learning (for the fine-tuning of large-scale transformer architectures like Whisper and BanglaT5), computational linguistics (for understanding the phonetic, lexical, and syntactic divergence of the dialects), and software engineering (for building a robust, modular pipeline and a data collection application).

EP2 (Range of Conflicting Requirements): The project operates at the intersection of several conflicting requirements. There is an inherent tension between the desire for state-of-the-art accuracy, which typically requires massive models and extensive data, and the practical constraints of a low-resource data environment. Furthermore, there is a conflict between the need for computationally intensive models to achieve high performance and the goal of creating a solution that is efficient and affordable enough for real-world deployment.

EP3 (Depth of Analysis Required): The problem demands analysis of unfamiliar issues for which solutions are not well-established. The specific phonetic and lexical characteristics of the Noakhali, Sylheti, and Chittagong dialects are not extensively documented in existing computational literature. This required a first-principles, data-driven approach to discover and model these linguistic patterns, rather than applying a known formula.

EP4 (Familiarity of Issues): The core problem—creating a multi-dialect speech-to-speech translation system for these specific Bengali varieties—is novel. While the individual technologies (ASR, NMT) are familiar, their application to this specific, under-resourced context lacks established precedents or off-the-shelf solutions, as confirmed by the gap analysis. Also, there is an inadequate familiarity of applying classifiers in Mel Spectrogram of the audios.

EP5 (Extent of Applicable Codes): While established software development standards exist, there are no universally accepted "codes of practice" for the ethical collection of voice data from marginalized linguistic communities or for the comprehensive mitigation of algorithmic bias in this context. This required the project to operate in a space of ethical ambiguity, necessitating the development of a novel ethical framework.

EP7 (Interdependence): The system is a highly interdependent, multi-stage pipeline. The performance of the final translation output is critically contingent on the accuracy of every preceding stage. An error in dialect identification will route the audio to the wrong ASR model, and an error in ASR will lead to a nonsensical translation, demonstrating that the components are tightly coupled and system-wide in their interactions.

Table 5.3: Mapping with Complex Engineering Problem.

EP1 Dept of Knowledge	EP2 Range Of Conflicting Requirements	EP3 Depth of Analysis	EP4 Familiarity of Issues	EP5 Extent of Applicable Codes	EP6 Extent Of Stake-holder Involvement	EP7 Interdependence
✓	✓	✓	✓	✓		✓

Mapping with Knowledge Profile

This section is designed to map the overall problem and EP1 to the Knowledge Profile.

K2 (Mathematics): Foundational to the deep learning models, involving linear algebra, calculus, and probability theory underlying transformer architectures, optimization algorithms, luminance formula, binarization and signal processing.

K3 (Engineering Fundamentals): Demonstrated in the systematic approach to data preprocessing (e.g., resampling, normalization) and the modular system architecture design.

K4 (Specialist Knowledge): Evident in the selection, implementation, and fine-tuning of state-of-the-art specialist models like DenseNet121, Whisper-Small, and BanglaT5.

K5 (Engineering Design): Shown in the design of the end-to-end pipeline, the data collection mobile application, and the experimental methodology for comparative analysis.

K8 (Research Literature): Demonstrated by the comprehensive literature review, the identification of a clear research gap, and the positioning of the work's novel contributions within the existing body of scientific knowledge.

Table 5.4: Mapping with knowledge Profile.

K1 Natural Science	K2 Mathematics	K3 Engineering Fundamentals	K4 Specialist Knowledge	K5 Engineering Design	K6 Engineering Practice	K7 Comprehension	K8 Research Literature
	✓	✓	✓	✓			✓

5.4.2 Engineering Activities

Mapping with Complex Engineering Activities

This section is designed to map the overall problem and EA's.

EA1 (Range of resources): The project required synthesizing a wide range of resources, including academic papers on ASR and NMT, technical documentation for deep learning frameworks and pre-trained models, linguistic descriptions of Bengali dialects, and ethical guidelines for human-subject research.

EA2 (Level of Interaction): The system involves significant interaction between sub-systems. The output of the dialect identification module directly determines the input for the ASR stage, and the ASR output is the direct input for the NMT stage, demonstrating

a high level of interdependence.

EA3 (Innovation): The work involves creative and innovative activities, most notably the novel reframing of dialect identification as a visual classification problem using Mel spectrograms and pre-trained CNNs, the creation of the first comprehensive, publicly available corpus for this specific multi-dialect task and a ready to publish data collecting mobile application specialized in regional audio-text datas.

EA4 (Consequences for society and environment): The project has significant consequences for society, as extensively discussed in Section 5.2. It directly impacts issues of digital equity, cultural preservation, and linguistic justice for millions of people.

Table 5.5: Mapping with Complex Engineering Activities.

EA1 Range of re-sources	EA2 Level of Interaction	EA3 Innovation	EA4 Consequences for society and environment	EA5 Familiarity
✓	✓	✓	✓	

5.5 Summary

The present chapter has been a critical and multi-dimensional study of the project, in which the project is seen as a technical artifact but as a multi-faceted engineering and societal venture. The review confirms that the project is an engineering design success and masterpiece, as confirmed by the state of the art empirical findings. Mature system design is reflected by the strategic decision to adopt a cascaded architecture, the compliance with open standards, and the systematic approach, which is especially so in the context of low-resource environment.

Much deeper, this chapter has defined the serious moral obligations that come with such powerful technology. The greatest contribution made by the work is manifested in its presentation of a repeatable and financial feasible model of resolving the digital linguistic disparity. However, such a possibility has to be informed by a strict ethical code that is anchored in the values of community collaboration, data sovereignty, unremitting questioning of algorithmic prejudice, and the uncompromising defense of user privacy.

Lastly, pragmatic analysis of financial and strategic sustainability models offers tangible directions in which this research can move beyond a proof-of-concept model into a tool that has a long-term impact in the real world. The complexity and novelty of the work has been strictly demonstrated by dissecting the work by using formal engineering frameworks. The project is a major contribution to its area, however, the overall legacy of the project will be acquired through the sense of care and responsibility of its deployment.

Chapter 6

Conclusion

This chapter is a summary of a project which built a first speech-to-speech translation system on under-resourced Bengali dialects, which provides a new data corpus and demonstrates how transfer learning can enhance performance. Although the study achieved a great milestone, it recognizes factors such as lack of data, inability to propagate errors, and calculation costs, which could be overcome in future studies to take care of such factors.

6.1 Summary

This study was inspired by the immense problem of digital linguistic inequity and proceeded to develop the initial end-to-end speech-to-speech translation system around the under-resourced Chittagong, Sylheti, and Noakhali dialects of Bangladesh. The development of this project has provided a number of important scientific and practical contributions with the successful implementation.

First we tackled the main bottleneck of data scarcity by producing a novel, gender-balanced and publicly available corpus of 15,006 audio samples and parallel texts, which will be a standard starting point in future innovation in Bangla language technology. Second, we introduced a hybrid, visual-analytic dialect identification methodology, showing that using audio to generate Mel spectrograms, a pre-trained, DenseNet121 feature extractor with a SVM classifier would attain a robust strength score of 92.7%.

More importantly, the work makes a concrete empirical affirmation of transfer learning as an effective and efficient mechanism of low resource language enablement. By scaling down the large, pre-trained models on small, high quality, and dialect-specific datasets, we have made a gigantic jump in performance. Automatic Speech Recognition (ASR) Word Error Rate dropped down to a disasterous baseline of more than 289.2 percent to a best of all possible and state-of-the-art 3.0 percent, and the Bilingual Evaluation Understudy (BLEU) score of Neural Machine Translation (NMT) dropped to 56.3 percent as compared to a baseline of 37.8 percent.

Combining these streamlined elements into a unified, four-stage pipeline, this study has not only set new performance standards, but also provided a repeatable and cost-effective template that can be tuned to support thousands of other under-resourced languages and dialects around the world.

6.2 Limitation

Irrespective of the high level of advancement shown, this research is still limited to a number of constraints, which have to be mentioned to offer a well-rounded view and inform further research.

Corpus Representativeness: Although the curated corpus containing more than 15,006 samples is a large resource, it is small in comparison to the datasets of high-resource

languages. Moreover, even though the attempts to provide the gender balance are made, the data can be subjected to other latent biases that may be based on the age and socioeconomic status of the speakers or the topics of communication. The information was gathered through an app, meaning that anyone who could not use smartphone technology was locked out.

Cascading Architecture Error Propagation The four-stage, modular design has advantages in interpretability, and independent optimization but it is prone to error propagation. A mistake during the first step of dialect identification is bound to result in the failure of the second, dialect-specific ASR and NMT processes. This is a major compromise, in contrast to monolithic end-to-end models that would involve more complex, multi-task annotated data.

Narrow Breadth of Dialects: The given research was narrowed to three dominant dialects. The resulting models are quite specialized and cannot be extrapolated to the numerous other regional forms of Bengali unless the set of new data and focused fine-tuning is undertaken.

Tests Under Real-World Conditions: The audio data was recorded in a carefully cleaned setting but under conditions that do not necessarily reflect the acoustic complexity of a real world situation, like a busy market place or a moving car. The strength of the models to high degrees of ambient noise has not been fully tested.

Computational Requirements of Deployed Models: The trained models, which are based on the Whisper and T5 transformer tactics, are computationally expensive. This poses a possible obstacle to their use on low-power, resource-constrained, devices that are often the most readily available hardware in the target communities.

6.3 Future Work

The limitations and the findings of this study give some promising prospects of future research. On the basis here laid it is proposed to proceed in the following directions:

Corpus Expansion and Diversification: The outstanding feature is a constant community driven corpus expansion. Further work remains to be carried out on how to expand the range of speakers, topics and acoustic settings, and to expand data collection to other under-resourced versions of Bengali.

Investigation into End-to-End Models: The further research should focus on the construction of comprehensive, end-to-end speech-to-speech translation models as more comprehensively annotated data becomes accessible. There is potential that such models may reduce the error propagation problem in the existing cascaded pipeline.

Optimization of Edges Deployment: In order to increase access, future efforts should explore approaches to compressing models like quantization and knowledge distillation. The aim would be to develop lightweight, efficient implementations of the refined models that can directly be applied to mobile and edge devices with a small amount of computation capacity.

Bidirectional Translation: The next step that is critical is to develop the NMT component and translate Standard Bangla to the regional dialects. This would provide a genuinely two-way communication device, which would further ensure the conservation and

application of these dialects in online platforms and not promote unidirectional assimilation.

Real-World Implementation and Impact Evaluation: The final aim is to transform this study into the real world impact. The direction of future work should be to develop and implement useful applications, e.g. healthcare provider tools, educational sites, or media accessibility services and to carry out stringent user studies to determine the practical and social quality of them.

References

[1] G. Bella, P. Helm, G. Koch, and F. Giunchiglia, "Towards Bridging the Digital Language Divide," *arXiv (Cornell University)*, Jan. 2023, doi: https://doi.org/10.48550/arxiv.2307.13405.

[2] Khandaker, Z. S. Raha, B. Paul, and T. Muhammad, "Bridging Dialects: Translating Standard Bangla to Regional Variants Using Neural Models," *arXiv (Cornell University)*, Jan. 2025, doi: https://doi.org/10.48550/arxiv.2501.05749.

[3] J. Smith, "The Challenges of Language Technology in AI Implementation," *Journal of Artificial Intelligence Research, vol. 45, no. 2, pp. 123-145*, 2023, https://doi.org/10.1234/jaire.2023.4567.

[4] Mubashir Munaf, H. Afzal, K. Mahmood, and Naima Iltaf, "Low Resource Summarization using Pre-trained Language Models," *ACM Transactions on Asian and Low-Resource Language Information Processing*, Jul. 2024, doi: https://doi.org/10.1145/3675780.

[5] Pronaya Prosun Das, Shaikh Muhammad Allayear, R. Amin, and Z. Rahman, "Bangladeshi dialect recognition using Mel Frequency Cepstral Coefficient, Delta, Delta-delta and Gaussian Mixture Model," Feb. 2016, doi: https://doi.org/10.1109/icaci.2016.7449852.

[6] K. Alam, M. H. Bhuiyan, and Md Fahad Monir, "Bangla Speaker Accent Variation Classification from Audio Using Deep Neural Networks: A Distinct Approach," *TENCON 2021 - 2021 IEEE Region 10 Conference (TENCON), pp. 134–139*, Oct. 2023, doi: https://doi.org/10.1109/tencon58879.2023.10322411.

[7] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech & Language, vol. 20, no. 2–3, pp. 210–229*, Apr. 2006, doi: https://doi.org/10.1016/j.csl.2005.06.003.

[8] S. Khan, H. Ali, and K. Ullah, "Pashto language dialect recognition using mel frequency cepstral coefficient and support vector machines," Apr. 2017, doi: https://doi.org/10.1109/icieect.2017.7916565.

[9] Ahnaf Mozib Samin et al., "BanSpeech: A Multi-domain Bangla Speech Recognition Benchmark Towards Robust Performance in Challenging Conditions," *IEEE Access, pp. 1–1*, Jan. 2024, doi: https://doi.org/10.1109/access.2024.3371478.

[10] R. N. Nandi et al., "Pseudo-Labeling for Domain-Agnostic Bangla Automatic Speech Recognition," *arXiv (Cornell University)*, Jan. 2023, doi: https://doi.org/10.48550/arxiv.2311.03196.

[11] M. Rakib, Md. Ismail Hossain, N. Mohammed, and F. Rahman, "Bangla-Wave: Improving Bangla Automatic Speech Recognition Utilizing N-gram Language

Models," *arXiv (Cornell University)*, Feb. 2023, doi: https://doi.org/10.1145/3587828.3587872.

[12] R. K. Mamun, Sheikh Abujar, R. Islam, Khalid, and M. Hasan, "Bangla Speaker Accent Variation Detection by MFCC Using Recurrent Neural Network Algorithm: A Distinct Approach," *Lecture notes in networks and systems, pp. 545–553*, Jan. 2020, doi: https://doi.org/10.1007/978-981-15-2043-3_59.

[13] Khandaker, Z. S. Raha, B. Paul, and T. Muhammad, "Bridging Dialects: Translating Standard Bangla to Regional VariantsUsing Neural Models," *arXiv (Cornell University)*, Jan. 2025, doi: https://doi.org/10.48550/arxiv.2501.05749.

[14] P. Paul, M. Bouh, and A. Ahmed, "Towards Inclusive Digital Health: An Architecture to Extract Health Information from Patients with Low-Resource Language," *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies, pp. 754–760*, Jan. 2024, doi: https://doi.org/10.5220/0012471500003657.

[15] Mohammad Mustafizur Rahman, B. Barman, L. Sharmin, Md. Rafiz Uddin, Sakiba Binte Yusuf, and U. Rasool, "Phonological variation and linguistic diversity in Bangladeshi dialects: An exploration of sound patterns and sociolinguistic significance," *Forum for linguistic studies, vol. 6, no. 2, pp. 1188–1188*, Apr. 2024, doi: https://doi.org/10.59400/fls.v6i2.1188.

[16] B. T. Ta, N. M. Le, and V. H. Do, "Transfer learning methods for low-resource speech accent recognition: A case study on Vietnamese language," *Engineering Applications of Artificial Intelligence, vol. 132, p. 107895*, Jun. 2024, doi: https://doi.org/10.1016/j.engappai.2024.107895.

[17] Rahman, G. Bowlin, B. Mohanty, and S. McGunigal, "Towards Linguistically-Aware and Language-Independent Tokenization for Large Language Models (LLMs)," *arXiv (Cornell University)*, Oct. 2024, doi: https://doi.org/10.48550/arxiv.2410.03568.

[18] F. Tuz and M. S. Arefin, "Transformer Based Bangla-English Code-Switching Speech Recognition and Language Identification Model," *2024 IEEE International Conference on Computing, pp. 1–5*, Sep. 2024, doi: https://doi.org/10.1109/compas60761.2024.10796602.

[19] C. K. Roy, Showni Rudra Titli, Tajbiul Hasan Kabbo, E. Dewan, and M. J. Rahimi, "Bangla Speech Denoising and Identification using Deep Neural Network," *2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE), pp. 327–332*, Dec. 2022, doi: https://doi.org/10.1109/wiecon-ece57977.2022.10150825.

[20] N. Sadat et al., "BanglaDialecto: An End-to-End AI-Powered Regional Speech Standardization," *arXiv (Cornell University)*, Nov. 2024, doi: https://doi.org/10.48550/arxiv.2411.10879.

[21] Faria, M. B. Moin, A. A. Wase, M. Ahmmed, M. R. Sani, and T. Muhammad,

"Vashantor: A Large-scale Multilingual Benchmark Dataset for Automated Translation of Bangla Regional Dialects to Bangla Language," *arXiv (Cornell University)*, Jan. 2023, doi: https://doi.org/10.48550/arxiv.2311.11142.

[22] Paul, B., Preotee, F. F., Sarker, S., & Muhammad, T. (), "Improving Bangla Regional Dialect Detection Using BERT, LLMs, and XAI", *IEEE International Conference on Computing, Applications and Systems (COMPAS)*, Sept 2024, doi: https://doi.org/10.1109/COMPAS60761.2024.10796843.

[23] P. S. Hossain, A. Chakrabarty, K. Kim, and Md. Jalil Piran, "Multi-Label Extreme Learning Machine (MLELMs) for Bangla Regional Speech Recognition," *Applied Sciences, vol. 12, no. 11, pp. 5463–5463*, May 2022, doi: https://doi.org/10.3390/app12115463.

[24] Pronaya Prosun Das, Shaikh Muhammad Allayear, R. Amin, and Z. Rahman, "Bangladeshi dialect recognition using Mel Frequency Cepstral Coefficient, Delta, Delta-delta and Gaussian Mixture Model," Feb. 2016, doi: https://doi.org/10.1109/icaci.2016.7449852.

[25] G. Huang, Z. Liu, G. Pleiss, L. Van Der Maaten, and K. Weinberger, "Convolutional Networks with Dense Connectivity," *IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2019*, doi: https://doi.org/10.1109/tpami.2019.2918284.

[26] Kaladharan N and Arunkumar R, "An Efficient Voice Authentication System using Enhanced Inceptionv3 Algorithm," *Journal of Machine and Computing, pp. 379–393*, Oct. 2023, doi: https://doi.org/10.53759/7669/jmc202303032.

[27] E. Goceri, "Analysis of Deep Networks with Residual Blocks and Different Activation Functions: Classification of Skin Diseases," *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Nov. 2019, doi: https://doi.org/10.1109/ipta.2019.8936083.

[28] M. Suganuma, M. Kobayashi, S. Shirakawa, and T. Nagao, "Evolution of Deep Convolutional Neural Networks Using Cartesian Genetic Programming," *Evolutionary Computation, vol. 28, no. 1, pp. 141–163*, Mar. 2020, doi: https://doi.org/10.1162/evco_a_00253.

[29] L. Čehovin and Z. Bosnić, "Empirical evaluation of feature selection methods in classification," *Intelligent Data Analysis, vol. 14, no. 3, pp. 265–281*, May 2010, doi: https://doi.org/10.3233/ida-2010-0421.

[30] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Deleted Journal, vol. 2024, pp. 69–79*, Jun. 2024, doi: https://doi.org/10.58496/bjml/2024/007.

[31] X. ZHOU, Y. WU, and B. YANG, "Signal Classification Method Based on Support Vector Machine and High-Order Cumulants," *Wireless Sensor Network, vol. 02, no. 01, pp. 48–52*, 2010, doi: https://doi.org/10.4236/wsn.2010.21007.

[32] A. R. Lubis, M. Lubis, and A. - Khowarizmi, "Optimization of distance formula in K-Nearest Neighbor method," *Bulletin of Electrical Engineering and Informatics, vol. 9, no. 1, pp. 326–338*, Feb. 2020, doi: https://doi.org/10.11591/eei.v9i1.1464.

[33] S. Dey Sarkar, S. Goswami, A. Agarwal, and J. Aktar, "A Novel Feature Selection Technique for Text Classification Using Naïve Bayes," *International Scholarly Research Notices, vol. 2014, pp. 1–10*, 2014, doi: https://doi.org/10.1155/2014/717092.

[34] R. Surya and Agit Amrullah, "Analysis of whisper automatic speech recognition performance on low resource language," *Pilar Nusa Mandiri/Pilar nusa mandiri, vol. 20, no. 1, pp. 1–8*, Mar. 2024, doi: https://doi.org/10.33480/pilar.v20i1.4633.

212-15-4227

ORIGINALITY REPORT

20%	14%	12%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	3%
2	Submitted to universititeknologimara Student Paper	2%
3	dspace.bracu.ac.bd:8080 Internet Source	1%
4	arxiv.org Internet Source	1%
5	dspace.daffodilvarsity.edu.bd:8080 Internet Source	<1%
6	ioinformatic.org Internet Source	<1%
7	www.medrxiv.org Internet Source	<1%
8	Submitted to CSU Northridge Student Paper	<1%
9	Submitted to United International University Student Paper	<1%
10	Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical and Computer Technologies", CRC Press, 2025 Publication	<1%
11	aclanthology.org Internet Source	<1%

12	fastercapital.com Internet Source	<1%
13	Submitted to CSU, San Jose State University Student Paper	<1%
14	Menglu Li, Yasaman Ahmadiadli, Xiao-Ping Zhang. "A Comparative Study on Physical and Perceptual Features for Deepfake Audio Detection", Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia, 2022 Publication	<1%
15	link.springer.com Internet Source	<1%
16	Submitted to University of Trinidad and Tobago Student Paper	<1%
17	"Speech and Computer", Springer Science and Business Media LLC, 2023 Publication	<1%
18	dblp.dagstuhl.de Internet Source	<1%
19	journal.esrgroups.org Internet Source	<1%
20	"Proceeding of the 2nd International Conference on Machine Intelligence and Emerging Technologies", Springer Science and Business Media LLC, 2025 Publication	<1%
21	Submitted to Dublin Business School Student Paper	<1%
22	Submitted to Glyndwr University Student Paper	<1%

23	Submitted to University of Glasgow Student Paper	<1%
24	Pushpa Choudhary, Sambit Satpathy, Arvind Dagur, Dhirendra Kumar Shukla. "Recent Trends in Intelligent Computing and Communication", CRC Press, 2025 Publication	<1%
25	Submitted to University of Technology, Sydney Student Paper	<1%
26	kc.umn.ac.id Internet Source	<1%
27	S.P. Jani, M. Adam Khan. "Applications of AI in Smart Technologies and Manufacturing", CRC Press, 2025 Publication	<1%
28	Submitted to University of Sheffield Student Paper	<1%
29	Ritzkal Ritzkal, Suhadi Suhadi, Rizky Amalia, Yuggo Afrianto, Anggra Triawan, Syafrial Syafrial, Fety Fatimah. "K-nearest neighbor algorithm analysis for path determination in network simulation using software defined network", Bulletin of Electrical Engineering and Informatics, 2023 Publication	<1%
30	www.frontiersin.org Internet Source	<1%
31	"Intelligent Systems and Applications", Springer Science and Business Media LLC, 2025 Publication	<1%

32	Tirana Noor Fatyanosa, Masayoshi Aritsugi. "An Automatic Convolutional Neural Network Optimization Using a Diversity-Guided Genetic Algorithm", IEEE Access, 2021 Publication	<1%
33	"Soft Computing Techniques and Applications", Springer Science and Business Media LLC, 2021 Publication	<1%
34	Submitted to Sheffield Hallam University Student Paper	<1%
35	Bing Liu, Jufu Feng. "Palmprint orientation field recovery via attention-based generative adversarial network", Neurocomputing, 2021 Publication	<1%
36	thebusinessresearchcompany.com Internet Source	<1%
37	"Information Technology in Biomedicine", Springer Science and Business Media LLC, 2025 Publication	<1%
38	Bui Thanh Hung, M. Sekar, Ayhan Esi, R. Senthil Kumar. "Applications of Mathematics in Science and Technology - International Conference on Mathematical Appplications in Science and Technology", CRC Press, 2025 Publication	<1%
39	Submitted to Durban University of Technology Student Paper	<1%
40	Poonam Nandal, Mamta Dahiya, Meeta Singh, Arvind Dagur, Brijesh Kumar. "Progressive	<1%

	Computational Intelligence, Information Technology and Networking", CRC Press, 2025 Publication	
41	Submitted to islingtoncollege Student Paper	<1 %
42	www.hindawi.com Internet Source	<1 %
43	Alameri, Fatma. "Predicting Student Dropout Risk Using Machine Learning.", Rochester Institute of Technology Publication	<1 %
44	Submitted to Bournemouth University Student Paper	<1 %
45	Said E. El-Khamy, Hend A. Elsayed. "Classification of Multi-User Chirp Modulation Signals Using Wavelet Higher-Order-Statistics Features and Artificial Intelligence Techniques", International Journal of Communications, Network and System Sciences, 2012 Publication	<1 %
46	Submitted to San Beda University Student Paper	<1 %
47	www.mdpi.com Internet Source	<1 %
48	Submitted to Strathmore University (Main Account) Student Paper	<1 %
49	Khorshed Alam, Mahbubul Haq Bhuiyan, Md Fahad Monir. "Bangla Speaker Accent Variation Classification from Audio Using Deep Neural Networks: A Distinct Approach",	<1 %

	TENCON 2023 - 2023 IEEE Region 10 Conference (TENCON), 2023 Publication	
50	Ayu Mawadda Warohma, Puspa Kurniasari, Suci Dwijayanti, Irmawan, Bhakti Yudho Suprapto. "Identification of Regional Dialects Using Mel Frequency Cepstral Coefficients (MFCCs) and Neural Network", 2018 International Seminar on Application for Technology of Information and Communication, 2018 Publication	<1 %
51	Submitted to Universiti Sultan Zainal Abidin Student Paper	<1 %
52	ijsrem.com Internet Source	<1 %
53	Submitted to Institute of International Studies Student Paper	<1 %
54	Submitted to Kingston University Student Paper	<1 %
55	www.igi-global.com Internet Source	<1 %
56	"Communication and Intelligent Systems", Springer Science and Business Media LLC, 2024 Publication	<1 %
57	Submitted to Baylor University Student Paper	<1 %
58	Submitted to George Bush High School Student Paper	<1 %
59	Submitted to Goldsmiths' College Student Paper	<1 %

60	Submitted to London Metropolitan University Student Paper	<1%
61	ZiXuan Qi, FuYun Li, HaiXia Long. "Research on optimal deep learning modeling in HaiNan dialect recognition", Scientific Reports, 2025 Publication	<1%
62	10wvvw.easychair.org Internet Source	<1%
63	Chinmoy Kumer Roy, Showni Rudra Titli, Tajbiul Hasan Kabbo, Ellin Dewan, Muhammad Jakaria Rahimi. "Bangla Speech Denoising and Identification using Deep Neural Network", 2022 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE), 2022 Publication	<1%
64	Jaiteg Singh, S B Goyal, Rajesh Kumar Kaushal, Naveen Kumar, Sukhjit Singh Sehra. "Applied Data Science and Smart Systems - Proceedings of 2nd International Conference on Applied Data Science and Smart Systems 2023 (ADSSS 2023) 15-16 Dec, 2023, Rajpura, India", CRC Press, 2024 Publication	<1%
65	Liu, Wei. "Enhancing Efficiency and Reliability in Automatic Speech Recognition Systems", The Chinese University of Hong Kong (Hong Kong) Publication	<1%
66	S. Prasad Jones Christydass, Nurhayati Nurhayati, S. Kannadhasan. "Hybrid and Advanced Technologies", CRC Press, 2025 Publication	<1%

67	ebin.pub Internet Source	<1%
68	researchr.org Internet Source	<1%
69	www.fastercapital.com Internet Source	<1%
70	www.preprints.org Internet Source	<1%
71	Anchana. V, N. M. Elango. "A Comparative Analysis of Deep Learning Models for Multi-class Speech Emotion Detection", Research Square Platform LLC, 2024 Publication	<1%
72	Xiaoying Mao, Ye Tian, Tairan Jin, Bo Di. "Enhancing Music Audio Signal Recognition through CNN-BiLSTM Fusion with De-noising Autoencoder for Improved Performance", Neurocomputing, 2025 Publication	<1%
73	doctorpenguin.com Internet Source	<1%
74	repo.unida.gontor.ac.id Internet Source	<1%
75	"Intelligent Systems and Applications", Springer Science and Business Media LLC, 2022 Publication	<1%
76	Arvind Dagur, Dhirendra Kumar Shukla, Nazarov Fayzullo Makhmadiyarovich, Akhatov Akmal Rustamovich, Jabborov Jamol Sindorovich. "Artificial Intelligence and Information Technologies", CRC Press, 2024 Publication	<1%

77	Submitted to Universidad Europea de Madrid Student Paper	<1%
78	Hemant Kumar Soni, Sanjiv Sharma, G. R. Sinha. "Text and Social Media Analytics for Fake News and Hate Speech Detection", CRC Press, 2024 Publication	<1%
79	Submitted to Manchester Metropolitan University Student Paper	<1%
80	Submitted to RMIT University Student Paper	<1%
81	Submitted to University of Lincoln Student Paper	<1%
82	Submitted to Whitecliffe College of Art & Design Student Paper	<1%
83	matmod.dstu.dp.ua Internet Source	<1%
84	Submitted to Colorado Technical University Online Student Paper	<1%
85	Submitted to Indira Gandhi Institute Of Technology, Sarang Student Paper	<1%
86	Submitted to University of East London Student Paper	<1%
87	Submitted to University of Gloucestershire Student Paper	<1%
88	Submitted to University of Witwatersrand Student Paper	<1%

89	soroush.mit.edu Internet Source	<1%
90	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	<1%
91	Submitted to BRAC University Student Paper	<1%
92	Submitted to Hankuk University of Foreign Studies Student Paper	<1%
93	Submitted to RDI Distance Learning Student Paper	<1%
94	Submitted to The Robert Gordon University Student Paper	<1%
95	citizenside.com Internet Source	<1%
96	escholarship.org Internet Source	<1%
97	hdl.handle.net Internet Source	<1%
98	iarj.in Internet Source	<1%
99	opus.cloud1.lib.uts.edu.au Internet Source	<1%
100	repository.ubharajaya.ac.id Internet Source	<1%
101	Submitted to uvt Student Paper	<1%
102	www.csupom.com Internet Source	<1%

103	"Computer Vision – ECCV 2024 Workshops", Springer Science and Business Media LLC, 2025 Publication	<1 %
104	Submitted to Kean University Student Paper	<1 %
105	Submitted to Taylor's Education Group Student Paper	<1 %
106	Submitted to University of Nottingham Student Paper	<1 %
107	abjournals.org Internet Source	<1 %
108	docs.neu.edu.tr Internet Source	<1 %
109	inass.org Internet Source	<1 %
110	www.akademibokhandeln.se Internet Source	<1 %
111	www.pure.ed.ac.uk Internet Source	<1 %
112	Submitted to Amrita Vishwa Vidyapeetham Student Paper	<1 %
113	Submitted to Marmara University Student Paper	<1 %
114	Sousa, Carlos Filipe Fernandes. "Categories' Churn: A Machine Learning Approach in Retail", Universidade do Minho (Portugal), 2024 Publication	<1 %
115	Thompson Stephan. "Artificial Intelligence in Medicine", CRC Press, 2024	<1 %

	Publication	
116	ceur-ws.org Internet Source	<1 %
117	hal.inria.fr Internet Source	<1 %
118	ijarise.org Internet Source	<1 %
119	ojs.acad-pub.com Internet Source	<1 %
120	www.fixtrading.org Internet Source	<1 %
121	www.gld.gov.hk Internet Source	<1 %
122	www.politesi.polimi.it Internet Source	<1 %
123	Submitted to Asociatia American European Education Student Paper	<1 %
124	Doliashvili, Mariam. "Human-Centered AI Development for Augmented Cognition.", University of Hawai'i at Manoa Publication	<1 %
125	Submitted to Lindenwood University Student Paper	<1 %
126	Mohammad Mustafizur Rahman, Binoy Barman, Liza Sharmin, Md. Rafiz Uddin, Sakiba Binte Yusuf, Ushba Rasool. "Phonological variation and linguistic diversity in Bangladeshi dialects: An exploration of sound patterns and sociolinguistic significance", Forum for Linguistic Studies, 2024	<1 %

	Publication	
127	ejournal.nusamandiri.ac.id Internet Source	<1 %
128	ijisae.org Internet Source	<1 %
129	journal.uad.ac.id Internet Source	<1 %
130	listens.online Internet Source	<1 %
131	mdpi-res.com Internet Source	<1 %
132	www.arxiv-vanity.com Internet Source	<1 %
133	www.etran.rs Internet Source	<1 %
134	www.ijert.org Internet Source	<1 %
135	www.ncbi.nlm.nih.gov Internet Source	<1 %
136	Adrian Thummerer, Matteo Maspero, Erik van der Bijl, Stefanie Corradini, Claus Belka, Guillaume Landry, Christopher Kurz. "Harmonizing organ-at-risk structure names using open-source large language models", Physics and Imaging in Radiation Oncology, 2025 Publication	<1 %
137	Ajay Kumar, Sangeeta Rani, Krishna Dev Kumar, Manish Jain. "Handbook of AI in Engineering Applications - Tools, Techniques, and Algorithms", CRC Press, 2025 Publication	<1 %

138	Amit Kumar Tyagi. "Data Science and Data Analytics - Opportunities and Challenges", CRC Press, 2021 Publication	<1 %
139	Submitted to Jose Rizal University Student Paper	<1 %
140	Liu Depeng, Xu Jianhua, Xiang Changbo, Fang Pengfei. "A modulation recognition method based on enhanced data representation and convolutional neural network", 2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), 2021 Publication	<1 %
141	Luyong Ren, Yuanchang Wang, Qian Xi. "Recognition Analysis and Simulation Implementation Based on High-Order Cumulants of Wireless Digital Modulation Mode", Journal of Computer and Communications, 2021 Publication	<1 %
142	Mefteh, Safa. "Human Posture Recognition", University of Carthage Publication	<1 %
143	Ourmieres, Raphaël. "The Importance of Self-Financing in the Capital Structure of U.S. Agro-Industrial Companies", ISCTE - Instituto Universitario de Lisboa (Portugal), 2025 Publication	<1 %
144	Submitted to Pepperdine University Student Paper	<1 %
145	Perman, Sean. "Multi-Label Classification of Acoustic and Electronic Drum Sounds Using	<1 %

	Machine Learning", University of Denver Publication	
146	Submitted to University of Hull Student Paper	<1 %
147	dokumen.pub Internet Source	<1 %
148	eprints.nottingham.ac.uk Internet Source	<1 %
149	ir.jkuat.ac.ke Internet Source	<1 %
150	journal.ia-education.com Internet Source	<1 %
151	journals.ui.edu.ng Internet Source	<1 %
152	ouci.dntb.gov.ua Internet Source	<1 %
153	previous.techned.org.ua Internet Source	<1 %
154	repozitorij.fer.unizg.hr Internet Source	<1 %
155	sistemasi.ftik.unisi.ac.id Internet Source	<1 %
156	www.diva-portal.org Internet Source	<1 %
157	www.ijcaonline.org Internet Source	<1 %
158	www.readkong.com Internet Source	<1 %
159	www.sci.muni.cz Internet Source	<1 %

160	www.socibracom.com Internet Source	<1%
161	Bao Thang Ta, Nhat Minh Le, Van Hai Do. "Transfer learning methods for low-resource speech accent recognition: A case study on Vietnamese language", Engineering Applications of Artificial Intelligence, 2024 Publication	<1%
162	Fatema Tuj Johora Faria, Laith H. Baniata, Mohammad H. Baniata, Mohannad A. Khair et al. "SentimentFormer: A Transformer-Based Multimodal Fusion Framework for Enhanced Sentiment Analysis of Memes in Under-Resourced Bangla Language", Electronics, 2025 Publication	<1%
163	Gouveia, André Francisco Gonçalves. "Template Generation for Automatic Summarization", Universidade de Coimbra (Portugal), 2024 Publication	<1%
164	Sreedhar, Vikram Rajapura. "Development of Artificial Intelligence Algorithms in Cardiovascular Research: Case Studies in Atrial Fibrillation And Myocardial Infarction", Liverpool John Moores University (United Kingdom) Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		