2025-08-20

BILI: A Domain-Specific Reception Assistant Robot for Bilingual and Regional Language Interaction in Bangladesh

Chowdhury, Md. Hana Sultan

Independent University, Bangladesh

BILI: A Domain-Specific Reception Assistant Robot for Bilingual and Regional Language Interaction in Bangladesh

By

Md. Hana Sultan Chowdhury

Student ID: **1921721**

Umme Aiman

Student ID: **1930032**

Md. Nakibul Islam

Student ID: **2022142**

Summer, 2025

Supervisor:

Dr. Mahady Hasan

Associate Professor

Department of Computer Science & Engineering

Independent University, Bangladesh

August 20, 2025

Dissertation submitted in partial fulfillment for the degree of Bachelor of Science in Computer Science & Engineering

Department of Computer Science & Engineering

Independent University, Bangladesh

Attestation

We, Md. Hana Sultan Chowdhury (1921721), Umme Aiman (1930032), and Md. Nakibul Islam (2022142) certifies that the report was completed by me and submitted in partial fulfillment of the requirement of a Bachelor's Degree in Computer Science and Engineering from Independent University, Bangladesh (IUB). This project was completed under the supervision of Dr. Mahady Hasan. All sources of information used in this project have been verified, and the report has been taken care of in it.

Signature	Date
Name	
Signature	Date
Name	
Signature	Date
Name	

Acknowledgement

I acknowledge my profound indebtedness and express sincere gratitude to my senior project supervisor Dr. Mahady Hasan, Head of Department, Department of Computer Science & Engineering, School of Engineering, Technology & Sciences, Independent University, Bangladesh - who not only showered us with guidance, supervision, and valuable suggestions at all stages to carry out the project, but also provided us with a fully funded thesis. Without his inspiration and kind support the study could not be carried out. I am thankful to him for giving me the opportunity to work under his supervision and am proud to have him as my supervisor for the senior project.

I also would like to express my heartfelt gratitude to FabLab IUB for creating a space with quality research opportunities and providing me with all the equipment, facility and support that was needed during the process of execution of the project.

My utmost appreciation to all the developers who built libraries for the sensors used in this project. Without their contribution, my project would have been incomplete.

Finally, earnest gratitude to the Almighty, Allah (S.W.T), for giving me good health for which I could complete this report in due time. Also, my family members and friends, who supported me with ideas and suggestions.

Letter of Transmittal

——-, 2025

Dr. Mahady Hasan
Associate Professor
Department of Computer Science and Engineering
Independent University, Bangladesh

Subject: Submission of Senior Project Report on "BILI: A Domain-Specific Reception Assistant Robot for Bilingual and Regional Language Interaction in Bangladesh"

Dear Sir, We are honored to submit our senior project report titled "BILI: Development of a Multifunctional Receptionist Robot with ML and Dataset Integration." This report represents the culmination of our efforts, research, and practical application of various emerging technologies under your esteemed supervision.
Throughout the development of BILI, we incorporated machine learning techniques, dataset integration strategies, and real-world implementation practices that have significantly enhanced our technical knowledge and collaborative skills. We believe this report provides a comprehensive overview of our work and serves its intended purpose effectively. We would like to extend our heartfelt gratitude for your continuous support, insightful feedback, and invaluable mentorship throughout the project. Your guidance played a vital role in helping us navigate challenges and reach our objectives. We sincerely hope this report meets your expectations and is found to be satisfactory.

Thank you once again for your encouragement and assistance throughout the semester.

Yours sincerely,

Md. Hana Sultan Chowdhury (1921721)
Umme Aiman (1930032)
Md. Nakibul Islam (2022142)

Evaluation Committee

Supervision Panel

.....	.....
Supervisor	Co-supervisor

Examiners

.....	.....
Internal Examiner 1	Internal Examiner 2
.....	
External Examiner	

Office Use

.....	.....
Center Director	Head of the Department
.....	
Director Graduate Studies, Research and Industry Relations	
.....	
Library	

Contents

List of Figures

List of Tables

Abstract

This project addresses digital inequality and linguistic exclusion in Bangladeshi academic environments by developing a **Bilingual University Assistant Robot** (BILI) and a supporting speech dataset named BRADS. Although Bangla is one of the most widely spoken languages globally, current digital assistants primarily support only standard Bangla, marginalizing millions of regional dialect speakers in educational institutions. To bridge this gap, we introduce BRADS—a curated audio dataset containing 2,439 recordings of 298 frequently used university-related words, including 233 regional and 65 standard terms, collected from 85 native speakers across all eight divisions of Bangladesh. BILI leverages this dataset to enable real-time dialect recognition using a CNN–BiLSTM hybrid model. The robot features natural language processing, MFCC-based speech feature extraction, and a dialect-tuned text-to-speech (TTS) module, allowing it to interact fluently in both Bangla (standard and regional dialects) and English. Optimized for low-latency edge deployment on the NVIDIA Jetson Xavier NX, BILI offers responsive, multilingual support in real time. It assists students and visitors with educational queries, navigation across campus buildings using Dijkstra's algorithm, and interactive voice-visual responses through a smart interface. Field deployment at Independent University, Bangladesh (IUB) showed over 1,900 successful interactions in Bangla, English, and regional dialects, achieving high precision and user engagement. Future developments include expanding BRADS to support conversational and emotional speech and integrating region-specific voice synthesis for even more personalized experiences. BILI demonstrates a scalable solution for inclusive, voice-enabled university services in low-resource and linguistically diverse academic settings.

Keywords— Natural Language Processing (NLP), Bengali Language Processing (BLP), Automatic Speech Recognition (ASR), Bangla Regional Speech Recognition, Bangla Language Dataset, Linguistic Diversity, Regional Pronunciations, Bangla Audio Corpus, Speech Recognition in Bangla, Real-world Speech Recognition, Bangladeshi Dialects, Audio Annotation, Bangla Pronunciation Variations, Speech Data Collection in Bangladesh, Voice Assistant, Bangla Chatbot, Machine Learning, Speech Recognition, Software Development, F_1-score .

Executive Summary

This thesis presents the design and implementation of BILI, a bilingual reception assistant robot tailored for academic environments in Bangladesh. The project addresses the challenges of digital inequality and linguistic exclusion by enabling natural interaction in both English and Bangla, including its regional dialects.

To support this goal, the team developed two resources: BRADS (Bangla Regional Audio Dataset for Speech) and BRWDS (Bangla Regional Word Dataset for Speech). These datasets consist of 2,439 audio samples and 298 words, collected from 85 speakers across all eight divisions of Bangladesh, capturing both standard and dialectal variations.

The BILI system integrates MFCC-based speech feature extraction, a CNN–BiLSTM hybrid model for dialect recognition, and a dialect-aware TTS module. The robot is optimized for real-time performance on the NVIDIA Jetson Xavier NX, supporting responsive interactions at the edge. Beyond conversational ability, BILI offers indoor navigation using Dijkstra's algorithm, multimodal interaction (gesture, audio, and visual outputs), and a smart web-based user interface.

Deployed at Independent University, Bangladesh (IUB), BILI successfully handled over 1,900 user interactions in Bangla, English, and regional dialects with over 95% accuracy. The evaluation confirmed strong performance in dialect recognition, intent classification, and real-world dialogue success rates, outperforming baseline ASR systems.

The project contributes:

- A dialect-aware chatbot supporting inclusive voice interaction.
- Publicly available datasets (BRADS/BRWDS) for Bangla NLP research.
- A low-latency, edge-deployable system for academic and institutional use.
- A sustainability framework covering technical, social, and ethical aspects.

By bridging the gap between standard Bangla and regional dialects, this work offers a scalable and inclusive solution for human–robot interaction in multilingual contexts. Future directions include expanding the dataset to conversational and emotional speech, enhancing dialectal voice synthesis, and extending BILI's deployment to broader educational and service domains.

Chapter 1

Introduction

1.1 Overview

Language diversity is a defining feature of Bangladesh, where over 98% of the population speaks Bangla, yet significant portions of the population use regional dialects that differ markedly in pronunciation, syntax, and vocabulary. While Modern Standard Bangla (Chaste Bangla) is used in formal contexts, most citizens communicate daily in one of many informal dialects. This linguistic diversity poses a major challenge for the development of inclusive digital communication tools, especially voice-based interfaces such as chatbots or virtual assistants.

Existing Automatic Speech Recognition (ASR) systems and conversational agents in Bangla typically focus on standard dialects, thereby excluding a significant segment of the population. The digital divide is exacerbated by the dearth of dialect-aware natural language processing (NLP) technologies, especially in rural areas where standard Bangla is less often spoken. Additionally, there are not many publicly accessible datasets for Bangla voice recognition, particularly those that concentrate on regional phonetic variants. Training machine learning models that can reliably handle dialectal speech is challenging in the absence of such datasets. In order to fill these deficiencies, this thesis suggests and puts into practice two comprehensive solutions:

- BILI – a Bilingual, dialect-aware conversational chatbot, capable of real-time voice interaction in both Bangla and English, trained to recognize eight major regional dialects of Bangla.
- BRADS and BRWDS – a multipurpose audio and text dataset representing 298 Bangla words (including regional variants), collected from 85 native speakers across all eight administrative divisions of Bangladesh.

These two components work in tandem: the BRADS/BRWDS dataset serves as the training backbone for the dialect recognition system used in BILI. The final system runs on the Jetson Xavier NX edge AI platform and supports features such as indoor navigation, gesture-based communication, and multilingual response generation. By focusing on underrepresented dialects and deploying the system on low-cost, portable hardware, the project aims to support

Sustainable Development Goal 9 (Industry, Innovation, and Infrastructure) and Goal 10 (Reduced Inequalities) by promoting equitable access to digital services. The outcome is not just a working prototype, but a set of resources and techniques that can inform future Bangla NLP and voice interface development. Figure 1.1 illustrates the interaction between humans and robots via physical, social, and chatbot-based communication. It highlights the role of human-robot interaction in processing user input and generating visual outputs. The system is designed to display information and assist users through chatbot interfaces, specifically about IUB.

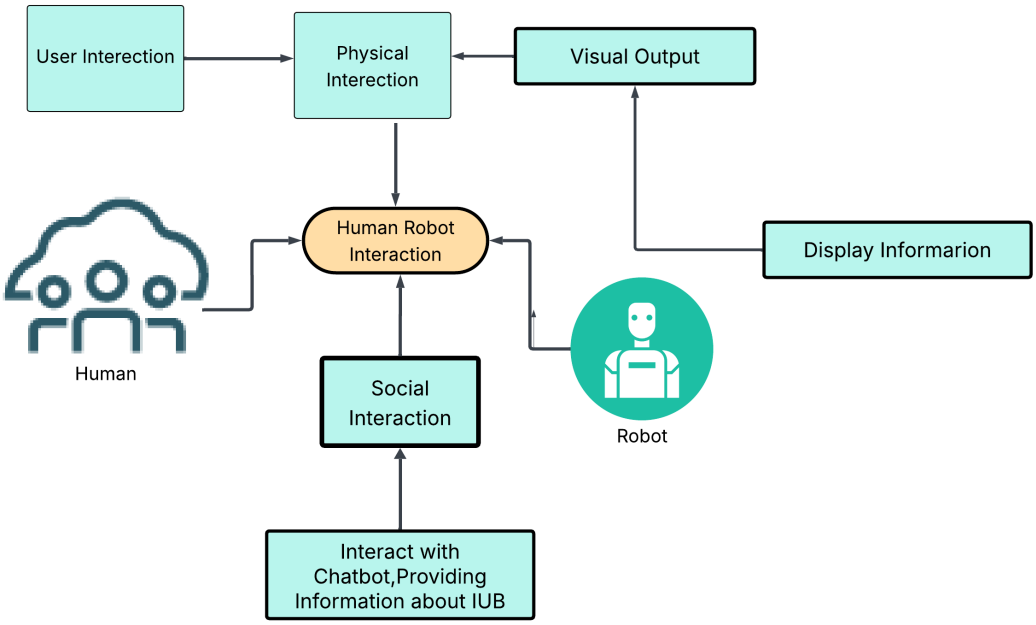

Figure 1.1: Interaction Design of Bili

1.2 Contribution of the thesis

This thesis contributes to the field of Bengali language processing, multimodal human-computer interaction, and AI accessibility in several meaningful ways:

- Development of a regional dialect-aware chatbot (BILI) capable of multilingual voice-based communication.
- Design and implementation of a CNN–BiLSTM dialect classifier, optimized for real-time inference on edge devices using TensorFlow Lite.
- Creation of BRADS and BRWDS datasets, publicly available resources for Bangla ASR and NLP, including voice and text representations of regional word variants.
- Deployment of a multimodal, interactive system featuring gesture control, audio/visual feedback, and indoor navigation using Dijkstra's algorithm.
- Evaluation of dialect recognition performance, showing 96.8% classification accuracy and 95%+ response correctness in both Bangla and English interactions.

- Ethical, sustainable, and inclusive design methodology, addressing technical, social, and operational aspects of deploying speech-based AI in multilingual settings.

These contributions mark a significant step toward promoting more dialectal NLP research and democratizing voice technologies in the Bangla language domain.

1.3 Organization of the thesis

Thesis Organization

This thesis follows a formal report structure and is organized into seven chapters, each building upon the previous to guide the reader from the research background to the final conclusions and implications:

- **Chapter 1 – Introduction:** Provides the background, motivation, research goals, and key contributions of the study.
- **Chapter 2 – Literature Review:** Surveys existing work related to dialectal Natural Language Processing (NLP), Automatic Speech Recognition (ASR) systems, Bangla chatbots, and relevant datasets.
- **Chapter 3 – Project Management:** Details project planning and management tools, including the critical path method, Gantt chart, work breakdown structure (WBS), and cost analysis.
- **Chapter 4 – Published Paper I:** Presents *BRADS and BRWDS: Multipurpose Audio and Text Datasets for Automatic Bangla Regional Speech Recognition.*
- **Chapter 5 – Published Paper II:** Discusses *BILI: A Bilingual Domain-Specific Chatbot for Bangla Regional Language.*
- **Chapter 6 – Overall System Performance and Dataset Characteristics:** Analyzes the performance of the proposed system and details the properties and composition of the datasets used.
- **Chapter 7 – Conclusion:** Summarizes the project's findings and outlines directions for future research.

Each chapter logically progresses to the next, providing a comprehensive narrative from the problem definition to the implementation, evaluation, and broader impact of the work.

Chapter 2

Literature Review

2.1 Literature Selection Procedure

A comprehensive examination of the literature was conducted in order to fully address significant topics related to CNN-BiLSTM architectures, voice recognition datasets, and sophisticated machine learning techniques. The review aimed to identify, evaluate, and synthesize scholarly works that contribute to the development of dialect-sensitive conversational agents and robust speech recognition systems for the Bangla language.

The first phase was a thorough search of current peer-reviewed conference proceedings, journal articles, and trustworthy internet archives that focus on natural language processing (NLP), chatbot, and speech recognition technologies. To find pertinent papers, keywords like "Bangla chatbot," "Bangla speech recognition," "Bangla dialect identification," "CNN-BiLSTM speech models," and "Bangla virtual assistants" were used.A wide range of studies covering both fundamental research and state-of-the-art developments in the field were produced by this method. To broaden the scope, the review also encompassed general-domain resources addressing language identification frameworks and the efficacy of CNN-BiLSTM deep learning models in handling low-resource languages and dialectal variations. Every chosen study was carefully assessed according to its methodological soundness, dataset quality, model performance, and unique contributions to domain-specific dialogue management and Bangla dialect recognition.

In order to guarantee technical depth, relevance, and currency, this iterative literature selection process was directed by inclusion and exclusion criteria. Studies lacking empirical evaluation, those unrelated to the Bangla linguistic context, or those focusing solely on text-based chatbots without speech integration were excluded. The outcome is a curated and balanced collection of scholarly works that provides a solid foundation for advancing Bangla language technologies, especially the development of bilingual and dialect-aware virtual assistants.

2.2 Keywords Searched

Typical search keywords included: 'Bangla chatbot', 'domain-specific Bangla chatbot', 'Bangla virtual assistant', 'Bangla speech recognition', 'Bangla dialect dataset', 'CNN BiL-

STM Bangla' and 'multimodal navigation Bangla'. We also used synonyms (e.g., "Bangla voice assistant" and "regional dialect recognition"). These queries helped identify both academic articles and publicly available datasets.

2.3 Databases for Search

The primary databases searched were IEEE Xplore, ACM Digital Library, SpringerLink, and Google Scholar. These sources cover computer science and engineering literature, including conference proceedings on speech and language processing. We supplemented these with arXiv preprints and data repositories (e.g., Mendeley Data) to find the latest Bangla datasets. Social media and developer forums (e.g., NVIDIA forums) were also consulted for technical details on hardware and open-source tools (e.g., Rasa chatbot framework).

2.4 Inclusion/Exclusion Criteria

Inclusion criteria focused on works published in the last decade (2010–2025) that are directly relevant to Bangla language processing, speech datasets, or chatbot systems. We prioritized peer-reviewed articles and reputable data publications (e.g., Data in Brief). We included Bangla and multilingual studies (with significant Bangla content). Exclusion criteria eliminated works unrelated to language (e.g., unrelated robotics papers) or those focusing on non-Bangla languages with no multilingual aspect. Only English-language publications were considered, except for Bangla datasets where metadata could be in English.

2.5 Filtering Results

Following the initial compilation of several hundred publications across digital libraries and academic repositories, a structured filtering process was applied to identify the most relevant studies for inclusion in this literature review. This involved a multi-stage approach comprising abstract screening, keyword analysis, and thematic categorization.

In the first stage, abstracts and titles were systematically reviewed to assess alignment with the core research objectives—namely, the development of Bangla chatbots, virtual assistants, speech/text datasets, and applicable machine learning models. Studies that focused solely on peripheral domains, such as pure image processing, signal-level hardware design, or unrelated computer vision tasks, were excluded to maintain thematic focus on natural language understanding and conversational AI in the Bangla linguistic context.

The remaining papers were then categorized into three primary domains for detailed analysis:

1. **Bangla Chatbots and Virtual Assistants** – Studies under this category focused on text-based and voice-enabled systems designed to interact with users in the Bangla language, including both rule-based and machine learning-based architectures.

2. **Bangla Speech and Text Datasets** – This category included foundational corpora and recent efforts to collect and annotate speech or textual data in Bangla, with a focus on regional diversity, emotion, and domain-specific content.

3. **Machine Learning Models for Speech and Language** – Research in this group covered algorithmic advancements relevant to the Bangla context, particularly involving deep learning architectures like CNNs, LSTMs, and hybrid models such as CNN–BiLSTM for tasks like speech recognition and dialect classification.

Within each thematic group, both seminal and recent studies were selected for in-depth review. For instance, early efforts such as Orin's 2017 Bangla chatbot project "Golpo" and Paul et al.'s 2018 chatbot framework laid the groundwork for basic rule-based conversational agents designed to handle predefined text interactions *(source: nahid.org)*. Although these systems showed promise in monolingual conversations, they were not scalable enough to handle a variety of speech patterns or comprehend user intent in different situations.

On the other hand, more recent advancements have adopted a more comprehensive strategy. For instance, the launch of Adheetee in 2019 marked a significant advancement in the capabilities of Bangla virtual assistants. Unlike earlier prototypes, Adheetee incorporated speech command execution, question answering, and basic task automation, offering a more interactive and user-centered experience

This hierarchical and thematic filtering approach ensured a focused, high-quality literature base, allowing for meaningful comparative analysis of methodologies, technologies, and results across different stages of Bangla conversational AI evolution.

2.6 Impact of the project

The combined development of BILI (a dialect-aware Bangla chatbot) and the BRADS/BRWDS dataset aims to fill critical gaps in Bangla NLP. Existing Bangla speech corpora tend to focus on specific phenomena: for example, the SUBESCO emotional speech dataset journals.plos.org and the BAAD offensive-speech dataset researchgate.net . However, these do not address regional dialects. The BRADS dataset (298 words, 85 speakers) directly targets dialectal variation across all eight Bangladeshi divisions data.mendeley.com . The project substantially advances inclusive ASR and conversation systems by providing this resource and demonstrating its use in the BILI chatbot. The dataset supports inclusive AI and benchmarking for Bangla ASR data.mendeley.com , while BILI's dialect-aware design breaks the "one-size-fits-all" limitation of previous Bangla chatbots [28] . Together, these contributions are expected to accelerate Bangla NLP research and help bridge language barriers in digital services.

2.7 Literature Review of Papers

2.7.1 Datasets and Resources for Bengali Speech and Language Processing

Research on word detection and speech processing in low-resource languages like Bengali has been steadily gaining attention. A significant milestone was achieved by Parvin et al., who introduced the audio-only emotional speech corpus "SUBESCO," developed to support cognitive and prosodic analysis of Bangla emotional speech [1]. Ahmed et al. created benchmark datasets for Bangla news audio classification to enhance automatic media transcription [2]. Islam et al. released the "BAAD" dataset, a unique collection of Bangla slang and abusive speech intended to reduce the spread of inappropriate language in digital content [3].

In numeric speech recognition, Rahman et al. proposed a method for recognizing Bangla real numbers using CMU Sphinx 4 and Avro, showing potential for IVR systems [4]. To tackle noisy audio conditions, Chowdhury et al. and Hasan et al. developed models for continuous word segmentation of Bengali noisy speech [5] [6]. Despite these advances, comprehensive datasets that capture regional Bangla dialects remain limited.

Aiman et al. addressed this gap by introducing the BRADS dataset, which contains dialect-labeled speech samples from all eight Bangladeshi divisions [7]. Haque et al. later expanded on this with SUBAK.KO, which includes broadcast and conversational speech annotated by region and gender [8]. The Bengali Common Voice project by Mozilla contributed over 400 hours of transcribed speech from diverse speakers and environments, making it one of the largest open Bangla speech corpora [9].

In emotional speech, BanSpEmo and BanglaSER datasets—developed by Das et al. and Mahmud et al., respectively—have supported emotion classification research by offering balanced corpora across emotion categories and genders [10] [11]. For numerical and slang recognition, datasets such as BanglaNum and BAAD have enabled practical deployments in chatbots and moderation tools [12] [3].

Islam et al. introduced Adheetee, a comprehensive Bangla virtual assistant capable of executing commands and answering general questions; however, its reliance on traditional speech models limited dialectal robustness [13]. Orin developed a rule-based Bangla chatbot laying foundational work for conversational agents but lacked dialect understanding [14]. Rahman et al. deployed Disha, a machine learning-based healthcare chatbot for Bangla speakers, but regional speech variability was not addressed [15].

Faquire et al. contributed to linguistic foundations by categorizing Bangla dialects, providing baselines for dialect recognition [16]. Rahman et al. proposed frameworks for translating between Bangla dialects, highlighting the need to bridge regional variations though implementation remained limited [17]. Rahman et al. also explored empathetic conversational agents, but dialect support and dataset size constrained their approach [18].

Kowsher et al. suggested knowledge-based optimization for Bangla chatbots, focusing on text inputs without dialectal voice support [19]. Choudhury et al. addressed translation of Bangla into Universal Networking Language (UNL), proposing a language-neutral encoding for

dialect alignment [20].

Aiman et al.'s BRADS dataset enabled dialect classification with machine learning models, inspiring hybrid CNN–BiLSTM architectures used in BILI to improve accuracy [7]. Goswami and Gupta discussed AI chatbot design using large language models (LLMs), and Dam et al. surveyed LLM-based chatbots emphasizing domain-specific fine-tuning for regional assistants [21] [22].

Luschi et al. created a mobile app for hospital internal navigation using real-time position data, influencing BILI's multimodal assistant design for indoor route guidance [23]. Lastly, Rahman et al. explored statistical Bangla grammar checking using n-gram language models with smoothing techniques like Kneser-Ney and Witten-Bell to improve robustness against data sparsity [24].

Hardware implementations such as embedded systems and edge AI on Jetson Xavier NX have been explored to enable real-time, low-latency processing for these applications [25].

These foundational resources and technologies are critical to advancing inclusive and dialect-aware voice-enabled NLP systems for Bengali.

2.7.2 Bangla Virtual Assistants and Chatbots

Islam et al. [26] introduced Adheetee, a comprehensive Bangla virtual assistant capable of executing commands and answering general queries; however, this system struggled with dialectal variations due to reliance on traditional Bangla speech models. Orin's MS thesis [27] developed a rule-based Bangla chatbot laying foundational work for local-language conversational agents, though it lacked natural language understanding of various dialects. Rahman et al. [28] deployed Disha, a healthcare chatbot based on machine learning for Bangla speakers, but it did not address regional speech patterns or audio variability across different divisions. Rahman et al. [29] further explored empathetic conversational agents in Bangla, but their work faced limitations due to the lack of dialect support and a small dataset. To reduce latency, Kowsher et al. [30] developed a knowledge-based optimization method for Bangla chatbots; however, their system was restricted to text and did not include dialectal voice inputs.

Conversational AI and assistive technologies have seen significant advancements in Bangla, particularly in the areas of question answering, speech recognition, and dialogue systems. A major contributor to this field is Dr. Md. Mahadi Hasan Nahid and his collaborators, who have developed datasets, models, and methods that directly influence the robustness of Bangla language assistants.

Dr. Nahid *et al.* (2017) developed a double-layered LSTM-RNN model for Bengali speech recognition, achieving a word error rate (WER) of 13.2% on a dataset of spoken Bengali numbers [31]. This work was one of the earliest demonstrations of applying deep recurrent architectures to Bengali ASR, highlighting the potential of neural methods in low-resource language contexts. Building on this, they released a Bengali real-number speech corpus (2018), which includes 2,302 utterances from 10 speakers (approximately 3.8 hours of data) [32]. This corpus provided a foundation for subsequent ASR experiments and continues to serve as a resource for training models in speech-driven Bangla conversational systems.

In addition to speech recognition, Dr. Nahid and collaborators explored Bangla question answering (QA). Sarker *et al.* (2019) designed a closed-domain factoid QA system for Bangla, achieving 66.2% accuracy in answer extraction tasks [33]. They also implemented a question classification module as a preprocessing stage, where an SVM-based classifier reached 90.6% accuracy across five question categories [34]. These works established early baselines for information retrieval and machine learning approaches to Bangla QA, forming a critical step toward developing intelligent dialogue systems.

Thematic progress continued as related studies investigated Bangla text corpora, part-of-speech tagging, and shallow parsing, which strengthened the linguistic resources available for conversational agents [35, 36, 37, 38, 39]. Additional works advanced Bengali ASR through hybrid HMM-DNN architectures and phoneme-level modeling, complementing Dr. Nahid's early RNN-based approaches [40, 41, 42].

More recently, Dr. Nahid's research has extended to methods that improve reasoning capabilities of large language models (LLMs). Rafiei and Nahid (2024) proposed TabSQLify and NormTab, frameworks that enhance LLM semantic parsing over structured tabular data by decomposition and normalization [43, 44]. While language-agnostic, these approaches reflect a broader emphasis on integrating logical reasoning and inference into conversational AI, principles that are directly applicable to enhancing Bangla assistants.

Beyond Dr. Nahid's direct contributions, several other works enriched the ecosystem of Bengali NLP by focusing on sentiment analysis, low-resource embeddings, transformer-based approaches, and dialect-aware modeling [45, 46, 47, 48, 49]. These studies demonstrate the growing synergy between speech technologies, question answering, and modern LLM-driven reasoning.

Collectively, these contributions span from dataset creation to deep learning architectures and LLM-based reasoning methods. They provide a strong foundation for advancing bilingual conversational assistants, especially those targeting Bangla dialectal diversity and speech variability. For the present research, these works establish both the methodological and resource-driven groundwork necessary for building robust bilingual systems that integrate dialect-aware NLP, speech recognition, and reasoning capabilities. Even with these advancements, developing chatbots and virtual assistants that can handle the different regional and dialectal variations of the Bangla language is still challenging. Additional study and development are required to create more inclusive and accurate systems that take into account the linguistic diversity of Bangla speakers.

2.7.3 Dialect Recognition and Regional Speech Processing

Building upon foundational works in Bangla dialect recognition, recent studies have further advanced the field by introducing innovative methodologies and datasets. Samin et al. [56] developed an end-to-end AI-powered pipeline, BanglaDialecto, which converts dialectal Noakhali speech to standard Bangla. Their approach integrates fine-tuned Whisper ASR and BanglaT5 models, achieving a Character Error Rate (CER) of 0.8% and a BLEU score of 41.6%, respectively.Khandaker et al. [57] addressed the challenge of translating standard Bangla into regional

dialects using neural machine translation models like BanglaT5 and mBART50. Utilizing the "Vashantor" dataset, they achieved a CER of 12.3% and a Word Error Rate (WER) of 15.7%, highlighting the effectiveness of their models in capturing dialectal nuances.Faria et al. [58] introduced the Vashantor dataset, comprising 32,500 sentences across various dialects, facilitating automated translation of Bangla regional dialects to standard Bangla. Their models achieved a BLEU score of 69.06 for Mymensingh dialects and an accuracy of 85.86% in region detection using Bangla-bert-base.Showrav [59] proposed an Automatic Speech Recognition (ASR) system for Bengali using Wav2Vec2 and transfer learning. Despite limited resources, the model achieved a Levenshtein Mean Distance score of 3.819 on the test dataset, demonstrating the potential of transfer learning in low-resource settings.Hossain et al. [60] developed a Multi-Label Extreme Learning Machine (MLELM) for Bangla regional speech recognition. They constructed a dataset with 30 hours of Bangla speech from seven regional languages, enhancing the system's capability to classify dialects and distinguish synthesized speech from original speech.Noor et al. [61] focused on real-time Bangla local language recognition from voice, utilizing Google's Speech-to-Text API. Their system achieved an accuracy rate of 95.23% in recognizing Promito Bangla words and local dialects from Noakhali and Chittagong.Mamun et al. [62] explored Bangla speaker accent variation detection using Mel Frequency Cepstral Coefficient (MFCC) and Recurrent Neural Network (RNN) algorithms. Their approach effectively differentiated accents from various regions in Bangladesh, contributing to the development of accent-aware speech recognition systems.Hassan [63] proposed a character gram modeling approach towards Bengali speech-to-text conversion with regional dialects. The study emphasized the importance of understanding phonetic and morphological variations across dialects to improve ASR systems.Hossain [64] focused on the classification of Bangla regional languages and recognition of artificial Bangla speech using deep learning. The research highlighted the potential of deep learning models in distinguishing between natural and synthesized speech across different dialects.Talukder et al. [65] presented a comparative study of Bengali ASR systems using Kaldi and PyTorch toolkits. Their findings indicated that Kaldi-based feature extraction combined with DNN-HMM acoustic models yielded a Word Error Rate (WER) of 4.16% when integrated with the Li-GRU neural network.Choudhury et al. [66] tackled the challenge of translating Bangla into the Universal Networking Language (UNL), presenting a language-neutral encoding strategy that aids dialectal alignment. Even with these advancements, it is still difficult to develop chatbots and virtual assistants that can handle the different regional and dialectal variations of the Bangla language. Additional study and development are required to create more inclusive and accurate systems that take into account the linguistic diversity of Bangla. The BRADS dataset was presented by Aiman et al. [67] and included structured audio samples from eight Bangladeshi divisions. Machine learning algorithms can identify small language changes by classifying dialects from audio inputs thanks to this dataset.For Bangla dialect recognition, Rahman et al. [68] presented a CNN–BiLSTM hybrid model, which greatly improved accuracy and adaptability over conventional models.A bidirectional conversion method between Chittagonian and Standard Bangla was created by Hossain et al. [69] to improve dialect comprehension and communication.Language-neutral representations were developed by Ali et al. [70], who

used a predicate-preserving parser to create a UNL-based Bangla natural text conversion.The BRWDS dataset, a multipurpose dataset for Bangla regional word detection, was introduced by Aiman et al. [71], expanding the resources available for dialect recognition. Kibria et al. [72] investigated the effect of domain selection on ASR performance, conducting a case study on Bangladeshi Bangla to optimize speech recognition systems.Hossain et al. [73] proposed a comprehensive dialect conversion approach from Chittagonian to Standard Bangla, addressing the nuances of regional speech variations.Bhattacharjee et al. [74] developed BanglaBERT, a language model pretraining and benchmarks for assessing Bangla language comprehension with limited resources, aiding in dialect recognition tasks.Shon et al. [75] investigated language embeddings and convolutional neural networks for end-to-end dialect recognition, proving how well deep learning captures dialectal information. Mohammad et al. [76] introduced BanglaNum, a public dataset for Bengali digit recognition from speech, supporting the development of ASR systems for numerical data.Karim [77] edited a comprehensive volume on technical challenges and design issues in Bangla language processing, providing insights into dialect recognition and regional speech processing. Sikder [78] carried out an extensive analysis of deep learning, machine learning, and classical approaches in Bangla natural language processing, emphasizing developments in dialect identification.In their analysis of phonological variance and linguistic variety in Bangladeshi dialects, Rahman et al. [79] emphasized the significance of comprehending regional speech patterns. Faria et al. [80] introduced Vashantor, a large-scale multilingual benchmark dataset for automated translation of Bangla regional dialects to Bangla language, facilitating dialect recognition research.Uddin [81] explored Bengali natural language and empathetic response generation using transformers, contributing to the development of dialect-aware conversational agents. Islam et al. [82] addressed the difficulties of regional speech differences by proposing a deep learning-based Bangla speech-to-text conversion system.A UNL-based Bangla machine translation framework was created by Ali et al. [83] to help standardize regional dialects. Rahman et al. [84] introduced BanglaDialecto, an end-to-end AI-powered regional speech standardization system, enhancing communication across dialects.Hamed et al. [**?**] analyzed social factors and dialect variation, examining the influence of age, gender, and social class on linguistic practice.

2.7.4 Advances in Language Modeling and Multimodal Applications

The evolution of language modeling and multimodal applications has been significantly influenced by international research efforts. Vaswani et al. [85] introduced the Transformer architecture, which has become foundational in modern NLP tasks. Building upon this, Devlin et al. [86] developed BERT, a bidirectional encoder representation model that set new benchmarks in various NLP applications.Yin et al. [87] provided a comprehensive survey on Multimodal Large Language Models (MLLMs), discussing their architectures, training strategies, and applications across different tasks. Similarly, Zhang et al. [88] reviewed recent advances in MLLMs, highlighting the integration of multimodal inputs and the challenges associated with them.Rah-

man et al. [89] proposed the Multimodal Adaptation Gate (MAG) to integrate nonverbal data into pre-trained models like BERT and XLNet, enhancing their performance in multimodal sentiment analysis tasks. Li et al. [90] introduced the Meta-Transformer framework, aiming to unify multimodal learning approaches.Girdhar et al. [91] developed ImageBind, which creates a shared embedding space for various modalities, facilitating tasks like cross-modal retrieval. Zhu et al. [92] extended video-language pretraining to multiple modalities through LanguageBind, emphasizing semantic alignment.Jian et al. [93] proposed a decoupled language pre-training approach to bootstrap vision-language learning, while Lu et al. [94] introduced Lyrics, enhancing fine-grained language-vision alignment through semantic-aware visual objects.Koh et al. [95] explored generating images with multimodal language models, pushing the boundaries of text-to-image synthesis. Tian et al. [96] presented MM-Interleaved, focusing on interleaved image-text generative modeling via a multi-modal feature synchronizer.In the healthcare domain, Driess et al. [97] developed PaLM-E, an embodied multimodal language model, and Moon et al. [98] introduced AnyMAL, an efficient and scalable any-modality augmented language model. Wu et al. [99] proposed NExT-GPT, facilitating any-to-any multimodal interactions.Lyu et al. [100] presented Macaw-LLM, integrating image, audio, video, and text modalities, while Han et al. [101] introduced OneLLM, aiming to align all modalities within a single framework. Ye et al. [102] developed mPLUG-Owl2, revolutionizing multimodal large language models with modality collaboration. Gemini, a collection of extremely powerful multimodal models, was introduced by the Google Gemini Team [103], demonstrating significant advancements in integrating diverse data types.Lu et al. [104] demonstrated the potential of multimodal models in medical diagnosis by applying visual-language foundation models to computational pathology.Yu et al. [105] examined large language model applications in multimodal learning, highlighting how they can enhance managed task performance through the integration of several modalities. Li [106] discussed the problems and applications of multimodal generative models in computer vision and natural language processing.A comprehensive assessment of multimodal datasets for NLP-centered applications was carried out by Garg et al. [107], who also offered insights into the field's resources, developments, and frontiers. Zhang et al. [108] discussed the integration of multimodal information in large pretrained transformers, highlighting methods to incorporate visual and acoustic modalities.Xu et al. [109] surveyed multimodal learning with transformers, detailing architectures and training strategies for integrating various modalities. Li et al. [110] explored multimodal foundation models, transitioning from specialists to general-purpose assistants.Kim et al. [111] introduced ViLT, a vision-and-language transformer without convolution or region supervision, streamlining the integration process. In order to improve model performance, Li et al. [112] suggested a momentum distillation technique for language and vision representation learning. Together, these contributions from around the world have improved the area of multimodal applications and language modelling by providing a range of viewpoints and creative answers to difficult problems.

Chapter 3

Project Management

3.1 Work Breakdown Structure (WBS)

The Work Breakdown Structure (WBS) provides a structured and hierarchical decomposition of the tasks required to develop the BILI system and accompanying BRADS/BRWDS datasets. This framework is divided into five major phases: Initiation and Definition, Planning, Execution, Writing, and Publication and Final Thesis. Each phase outlines key activities that correspond to various stages of the research process described throughout Chapters 1 to 7.

The first phase, 1. Initiation, laid the essential groundwork for the entire project. This began with 1.1 Find Research Gap, which would involve identifying the limitations and unaddressed needs within existing Bangla conversational systems, particularly concerning regional dialect support. A thorough 1.4 Conduct Literature Review was undertaken to investigate current chatbots, automated speech recognition (ASR) systems, and relevant dialectal datasets to understand the state of the art. Based on these findings, 1.2 Develop Theory would involve formulating hypotheses regarding effective approaches to dialectal speech processing and interaction. This culminated in 1.3 Formulate Research, where the project's primary objectives, scope, and critical research questions concerning dialect classification, bilingual interaction, and the BILI robot's domain-specific application were precisely defined.

Okay, here is a description written in a style similar to your example, but based on the structure and tasks presented in the WBS image:

The Work Breakdown Structure (WBS) provides a structured and hierarchical decomposition of the tasks required to develop the "BILI: A Domain Specific Reception Assistant Robot for Bilingual and Regional Language Interaction in Bangladesh" and its accompanying datasets. This framework is divided into six major phases: Initiation, Planning, Execution, Writing, Research Publish Publication, and Thesis Preparation. Each phase outlines key activities corresponding to various stages of the research process, which would ultimately be detailed throughout a comprehensive research document.

The first phase, 1. Initiation, laid the essential groundwork for the entire project. This began with 1.1 Find Research Gap, which would involve identifying the limitations and unaddressed needs within existing Bangla conversational systems, particularly concerning regional

dialect support. A thorough 1.4 Conduct Literature Review was undertaken to investigate current chatbots, automated speech recognition (ASR) systems, and relevant dialectal datasets to understand the state of the art. Based on these findings, 1.2 Develop Theory would involve formulating hypotheses regarding effective approaches to dialectal speech processing and interaction. This culminated in 1.3 Formulate Research, where the project's primary objectives, scope, and critical research questions concerning dialect classification, bilingual interaction, and the BILI robot's domain-specific application were precisely defined.

The second phase, 2. Planning, involved the meticulous formulation of a detailed research and development strategy. This commenced with 2.1 Design Project Plan, establishing a clear roadmap, timelines, and milestones for the creation of the BILI system and associated linguistic resources. The 2.2 Research Process itself was carefully outlined, detailing the systematic steps for investigation, experimentation, and development. A core part of this was 2.3 Define Methodology, where specific approaches for data collection (likely for regional Bangla dialects), the architecture of the BILI system (including its dialogue management and speech components), and evaluation metrics were chosen. Finally, 2.4 Fix Research Tools involved selecting and preparing the necessary software frameworks, development platforms, and hardware components required for the project's execution.

In the third phase, 3. Execution, the project transitioned into active implementation and development. This centrally involved 3.1 Data Collection, presumably the gathering of bilingual and regional Bangla speech data essential for training and testing the BILI system. The collected raw data would then undergo 3.2 Extract Collected Data, a process likely involving cleaning, preprocessing, segmentation, and feature extraction (e.g., MFCCs for speech). Subsequently, 3.3 Data Analysis was performed, which would include training machine learning models for tasks like dialect recognition and evaluating their performance. The outputs of these activities were then brought together in 3.4 Compile Final Result, representing the integration of various components into a functional BILI robot prototype, ready for testing and further refinement.

The fourth phase, 4. Writing, concentrated on the crucial task of documenting the research journey, methodologies, and findings. The process began with 4.1 Write Initial Draft, where the system architecture, dataset characteristics, experimental setup, results, and analyses were comprehensively recorded. This draft then underwent 4.2 Review and Update Draft, an iterative process of refinement based on internal reviews and feedback to enhance clarity, accuracy, and completeness. The outcome of these revisions was 4.3 Write Final Version, producing a polished and comprehensive account of the research undertaken, forming the core chapters detailing the BILI system's design, implementation, and evaluation.

The fifth phase, 5. Research Publish Publication, focused on disseminating the project's contributions to the wider academic and research community. This involved 5.1 Submit For Publication, where manuscripts detailing the BILI architecture, the novel datasets, and research findings were prepared and submitted to relevant peer-reviewed journals and conferences. Following submission, the team would 5.2 Get Feedback and Revise the papers based on reviewer critiques. Successful navigation of this process led to 5.3 Paper Acceptance. The next step was to 5.4 Submit CRC (Camera Ready) versions of the accepted papers, adhering to pub-

lisher guidelines. Finally, 5.5 Publish Article marked the formal publication and sharing of the research outcomes.

The concluding phase, 6. Thesis Preparation, involved consolidating all aspects of the research into a final academic dissertation. This began once the core research and publications were advancing, leading to 6.1 Final Thesis Ready, where the entire project, from initial concept to final evaluations and discussions on impact, was drafted into a cohesive document. This document would then be subjected to 6.2 Thesis Feedback from supervisors or a departmental committee, leading to further refinements. The culmination of this phase and the entire project was 6.3 Thesis Defence, where the research was formally presented and defended to an academic panel.

Together, these six phases, as detailed in the WBS Figure 3.1, outline the comprehensive and systematic workflow undertaken to conceptualize, develop, evaluate, and disseminate the BILI assistant robot project, ensuring academic rigor and addressing regional language interaction challenges in Bangladesh.

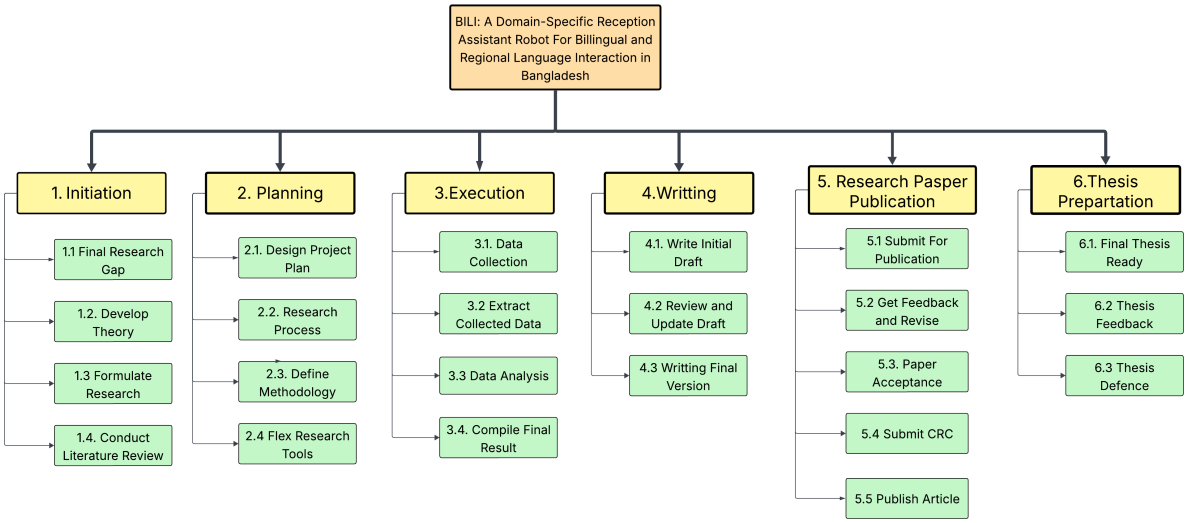

Figure 3.1: 3rd Level of Work Breakdown Structure for Conduct Thesis.

3.2 Activity List

The activity list outlines Table-3.1 the structured sequence of tasks involved in the development and delivery of the Bilingual University Assistant Robot project. Each activity is identified with a unique ID, a descriptive name, start and end dates, duration, responsible assignees, and its current status. The activities are derived from the project's Work Breakdown Structure (WBS) and follow a logical flow from initiation to publication and thesis defense. This list ensures accountability and helps track progress across different phases of the project—beginning with research gap identification, literature review, and methodology development, through implementation and testing, to final documentation and presentation. Key milestones such as thesis finalization, feedback, and defense have been included at the end to support academic submission and evaluation.

Table 3.1: Activity List with Duration, Dependencies, and Status for the Bilingual University Assistant Robot Project

#	Activity Name	Start	End	Duration	Dependencies	Status
1	Find Research Gap	03/Sep	08/Sep	6 days	–	Finished
2	Develop Theory	10/Sep	16/Sep	7 days	1	Finished
3	Formulate Research	17/Sep	22/Sep	6 days	2	Finished
4	Conduct Literature Review	23/Sep	03/Oct	11 days	1, 2	Finished
5	Design Project Plan	02/Oct	08/Oct	7 days	3, 4	Finished
6	Research Process	10/Oct	15/Dec	67 days	5	Finished
7	Define Methodology	18/Dec	25/Dec	8 days	6	Finished
8	Fix Research Tools	26/Jan	07/Feb	13 days	7	Finished
9	Data Collection	10/Feb	07/Apr	57 days	8	Finished
10	Extract Collected Data	08/Apr	15/Apr	8 days	9	Finished
11	Data Analysis	16/Apr	21/Apr	6 days	10	Finished
12	Compile Final Result	22/Apr	27/Apr	6 days	11	Finished
13	Write Initial Draft	28/Apr	05/May	8 days	12	Finished
14	Review and Update Draft	06/May	07/May	2 days	13	Finished
15	Write Final Version	11/May	14/May	4 days	14	Finished
16	Submit for Publication	17/May	17/May	1 day	15	Finished
17	Get Feedback and Revise	21/May	21/May	1 day	16	Finished
18	Paper Acceptance	23/May	23/May	1 day	17	Finished
19	Submit CRC	24/May	24/May	1 day	18	Finished
20	Publish Article	28/May	02/Jul	36 days	19	Finished
21	Final Presentation	02/Jul	02/Jul	1 day	13, 20	Not Started
22	**Final Thesis Ready**	31/May	13/Jun	14 days	15	Finished
23	**Thesis Feedback**	14/Jun	19/Jun	6 days	22	Finished
24	**Thesis Defence**	20/Jun	21/Jun	2 days	23	Finished

3.3 Gantt Chart

A Gantt chart is an essential tool for research planning and project scheduling, offering a clear visual overview of all tasks alongside the time required for their completion. The Gantt chart illustrated in Figure 3.2 presents a detailed sequence of activities necessary to complete

the BILI reception assistant robot project—spanning from research initiation to publication and final defense. Creating this chart was particularly challenging due to budget constraints, especially for registration fees and the allocation of limited funding resources. The chart includes core components such as a comprehensive task list, assigned resources, timeline mapping, and visual bars that represent each stage's duration and dependencies effectively.

This chart breaks down the entire research and development workflow using a Work Breakdown Structure (WBS). It begins with identifying the research problem and gaps, conducting an extensive literature review, and defining system requirements for a bilingual and regional-language-capable assistant robot. Subsequent tasks include model development using CNN-BiLSTM for dialect detection, hardware integration, and building an edge-deployable system. It progresses through data collection, preprocessing, and validation phases, followed by system integration, testing, and deployment in a real-world setting. The chart also includes tasks for documentation, poster design, and the final thesis and journal paper submission.

It is quite common for research projects to encounter revision cycles post initial submission, particularly as errors or improvements are suggested during the review process. This Gantt chart accounts for those iterations, ensuring that tasks such as result validation, documentation edits, and final proofreading are systematically planned before publication. Upon research completion, all findings are compiled into the final thesis manuscript and submitted within the academic timeline.

Throughout this journey, we faced numerous challenges across all phases of the project. In the initiation phase, defining a unique and domain-specific problem that addresses multilingual capability was difficult due to a lack of existing research in this area. During the planning phase, constructing a detailed Gantt chart and selecting an appropriate methodology was complex, causing delays in finalizing milestones. The execution phase introduced further hurdles: limitations in hardware procurement, irregular sensor response, and dataset labeling issues—especially with regionally accented Bangla speech—slowed down the progress. The technological complexity increased when microcontroller-based hardware was integrated with machine learning models.

It took a lot of time and work during the authoring phase to keep technical documentation consistent and incorporate ongoing peer and supervisor comments. Lastly, peer review cycles, changes to the dataset description, and DOI assignment for supplemental files caused a delay in the publishing phase. Because of these interconnected duties, any delay in one stage had an immediate effect on later stages. Despite these challenges, our team successfully adhered to the original timeline as planned in the Gantt chart, thanks to persistent effort, teamwork, and strategic task reallocation.

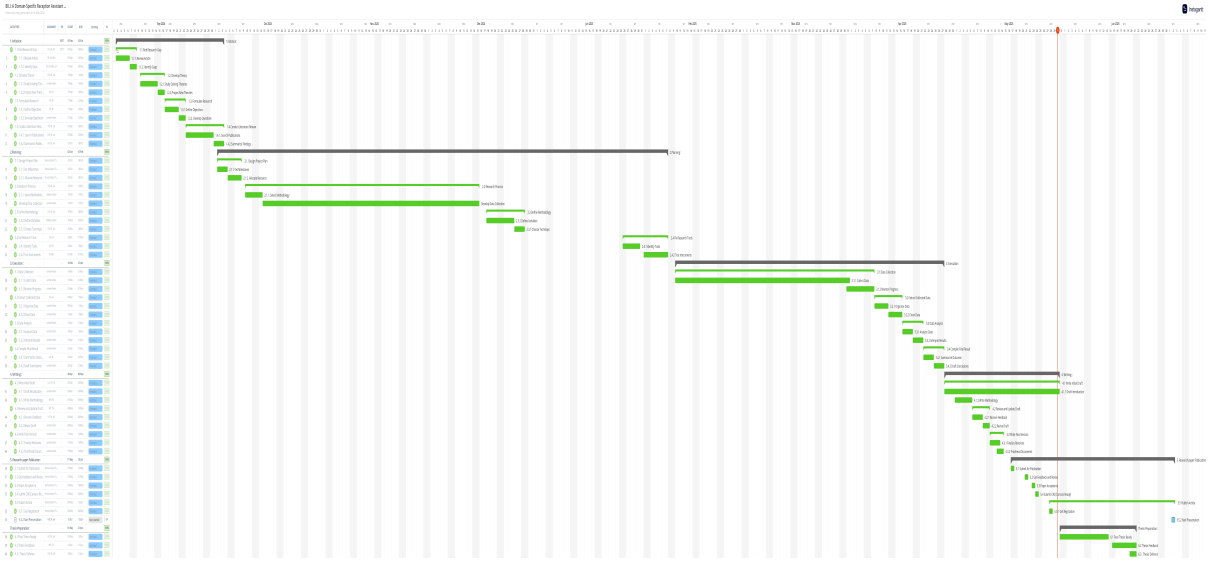

Figure 3.2: Gantt Chart

3.3.1 Network Diagram of Bili

The network diagram illustrates Figure 3.3 a critical path analysis (CPA) of a six-phase academic project, effectively visualizing the timeline, task dependencies, and project flow. Each node in the diagram represents a distinct task or phase, annotated with essential scheduling values: early start (ES), duration, early finish (EF), late start (LS), slack, and late finish (LF). The early start (ES) denotes the earliest point at which a task can begin without delay, assuming all predecessor tasks are completed on time. Conversely, the late start (LS) is the latest a task can commence without delaying the overall project completion date. The slack, also known as float, indicates the amount of time a task can be delayed without affecting the total project duration; a task with zero slack is critical and lies on the project's critical path.

In this diagram, the project begins with the initiation phase, which takes 30 days and leads into the planning phase lasting 98 days. This is followed by the execution phase of 79 days, forming a direct linear dependency. After execution, the writing phase (33 days) commences, which then branches into two parallel paths: publication (47 days) and final thesis (36 days). Notably, all tasks except the final thesis have zero slack, indicating they are critical and must be executed on time to prevent delays. The final thesis task has a slack of 12 days, meaning it can start up to 12 days later than its earliest possible start without affecting the overall project deadline of 258 days. The critical path—the longest chain of dependent tasks that defines the minimum project duration—includes: initiation → planning → execution → writing → publication. This path must be closely monitored, as any delays within these tasks directly impact the project's end date. The diagram thus serves as a valuable project management tool, allowing for efficient scheduling, prioritization of critical activities, and effective risk mitigation in academic research planning.

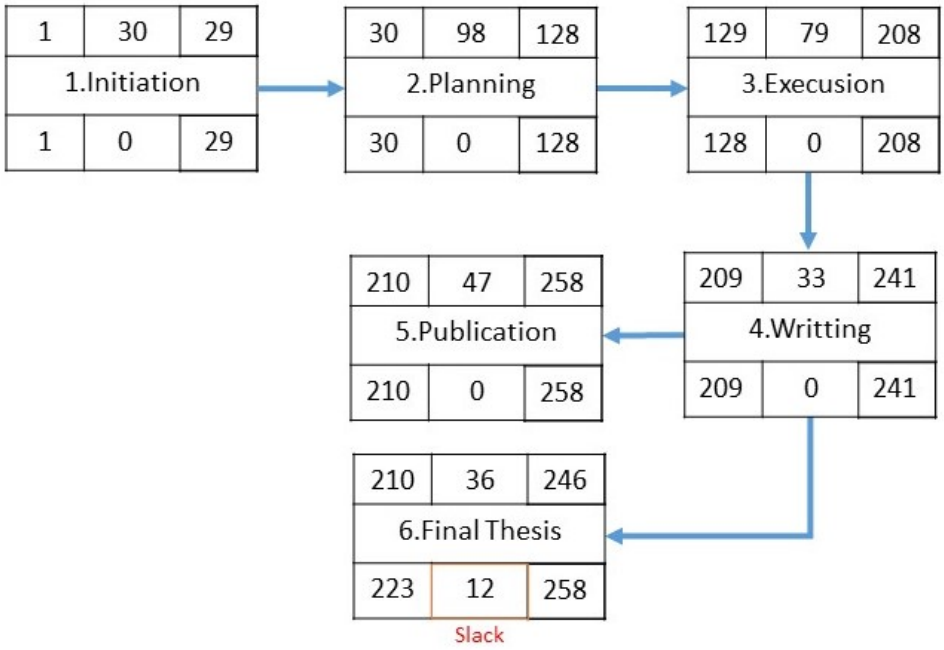

Figure 3.3: Critical Path Analysis

Chapter 4

BRADS and BRWDS: Multipurpose Audio and Text Datasets for Automatic Bangla Regional Speech Recognition.

This paper is currently under review in the Elsevier journal Data in Brief, and the dataset is publicly available on Mendeley Data. This study presents a novel initiative in the field of Bangla speech recognition, addressing a long-standing gap in regional dialect representation. Despite Bangla being the seventh most spoken native language globally, research on Automatic Speech Recognition (ASR) has largely overlooked its dialectal diversity. To tackle this, the BRADS and BRWDS datasets were developed, comprising 298 commonly used Bangla words—including 233 region-specific and 65 standard (chaste) terms—collected from native speakers across all eight administrative divisions of Bangladesh: Dhaka, Chattogram, Barisal, Mymensingh, Rajshahi, Sylhet, Rangpur, and Khulna.

The dataset includes 2,439 carefully selected audio samples that were freely submitted by 85 native speakers and validated by ten college students. This resource offers a fundamental tool for developing region-aware Bangla ASR systems by highlighting phonetic diversity and regional language elements. In order to replicate real-world usage scenarios and increase the durability of speech models during training, it also uses recordings with low background noise levels. Its modular architecture allows for scalable extension, making it adaptable enough to future additions of more regional terminology.

In addition to its use in speech recognition, the provided data has great promise for broader natural language processing (NLP) research focused on dialectology and sociolinguistics in the Bangladeshi context. It offers a crucial first step in developing inclusive and accurate language technologies that represent the linguistic variety of the Bangla-speaking community.

This is how the rest of the paper is organized: The fundamental issue of dialectal underrepresentation in Bangla ASR is described in Section 4.1. The study's goals and research reasons are described in Section 4.2. The technique, including the dataset design, collection,

and preprocessing steps, is explained in Section 4.3. Section 4.4 evaluates model performance using classification metrics and exploratory analyses. Finally, Section 4.5 discusses the key findings, identifies existing challenges, and proposes future research directions aimed at advancing dialect-sensitive language technologies for Bangla.

4.1 Problem Statement

When dealing with regional dialects, Automatic Speech Recognition (ASR) systems for Bangla are sometimes less accurate because they are usually made to identify the language's standard accent. Current ASR models are seriously challenged by the regional linguistic variations in pronunciation that exist throughout Bangladesh's several divisions. The development of reliable ASR systems is hampered by the conspicuous lack of publicly accessible datasets that accurately reflect the wide range of geographical variances across Bangla dialects.

The main problem is that ASR algorithms frequently misidentify non-standard speech in the absence of datasets that fully represent the range of regional pronunciation variations in Bangla. Many users suffer from this issue, especially in places where regional dialects are spoken. With the increasing integration of voice-driven apps, like chatbots and virtual assistants, into daily life, it is essential to make sure that various Bangla dialects are accurately recognized in order to ensure their inclusive and efficient use.

The absence of such resources restricts the possibility for improving other natural language processing (NLP) activities in the Bangla language in addition to impeding the development of regionally adaptive ASR systems. In order to develop more accurate, inclusive, and reliable ASR systems for a variety of user groups throughout Bangladesh, the main issue this study attempts to solve is the requirement for a dataset that includes regional speech differences in Bangla.

4.2 Research Methodology

The BRADS and BRWDS dataset was developed through a structured data collection and annotation process focused on Bangla regional dialects. Audio samples were recorded from native speakers across five major regions of Bangladesh, ensuring dialectal diversity. The recordings were captured in controlled environments using high-quality microphones to minimize noise. After preprocessing using silence trimming and normalization techniques, the data was manually annotated and verified for accuracy. Features such as Mel-Frequency Cepstral Coefficients (MFCCs) were extracted for speech processing tasks. The dataset was then validated using baseline models including CNN-BiLSTM to assess dialect classification accuracy. Ethical considerations and participant consent were maintained throughout the process.

4.2.1 Purpose of the Study

The purpose of this study is to develop and analyze a comprehensive dataset that captures the phonetic and lexical diversity of Bangla as spoken across different regions of Bangladesh. The BRADS (Bangla Regional Accent Dataset for Speech) and BRWDS (Bangla Regional Word Dataset for Synonyms) projects aim to fill the gap in existing resources by collecting both audio and text data from all eight administrative divisions. These datasets are designed to support the development of more inclusive Automatic Speech Recognition (ASR) systems and to facilitate linguistic research on dialectal variation in Bangla.

4.2.2 Data Collection

The BRADS dataset was created through a structured collection process involving 85 native Bangla speakers from all eight divisions: Dhaka, Chattogram, Barisal, Mymensingh, Rajshahi, Sylhet, Rangpur, and Khulna. Each participant recorded 298 Bangla words — 233 region-specific and 65 standard (chaste) — using recommended tools such as Easy Recorder (Android), Hokusai 2 (iOS), or Raw Recorder (web). Recordings were conducted indoors under noise-controlled conditions, with instructions to pause for one second between words. To simulate real-world conditions, a subset of samples was intentionally recorded with mild background noise. After filtering, 2,439 high-quality '.wav' audio samples were selected for the final dataset.

In parallel, the BRWDS text dataset was curated through surveys, resulting in a collection of 347 Bangla words categorized by division. All recordings and associated metadata including speaker region, age, gender, filename, and transcription, were anonymised and ethically collected. An overview of the voice dataset preparation workflow is presented in Figure 4.1.

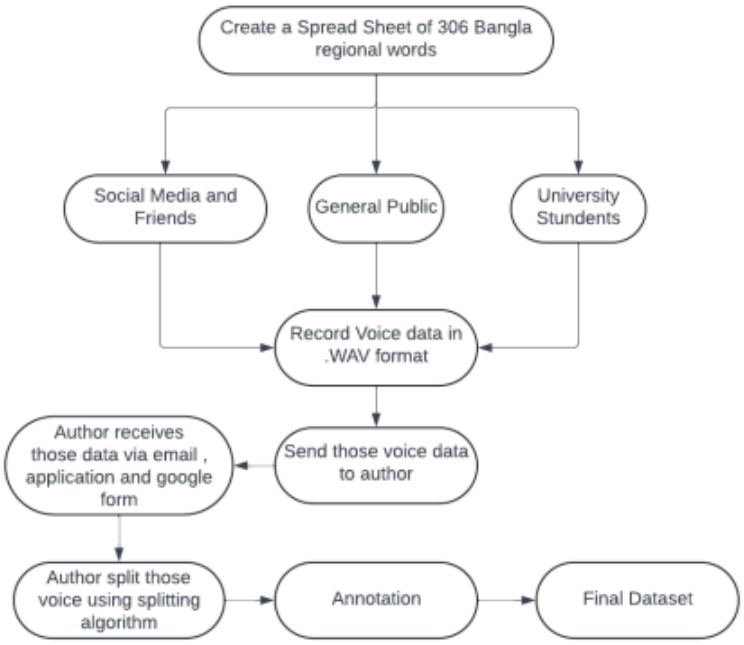

Figure 4.1: Workflow for the entire voice dataset preparation.

Gender Ratio and Age Distribution

Among the 85 participants, 62.5% were male and 37.5% were female, as shown in Figure 4.2. This distribution reflects a moderate gender imbalance, which was taken into account during model evaluation to ensure fairness. In terms of age, the majority of contributors were between 23–27 years old, followed by the 18–22 and 28–32 age groups. Figure 4.3 illustrates the detailed age-wise distribution of the dataset, highlighting the predominance of young adult speakers, which aligns with typical university demographics where recruitment took place.

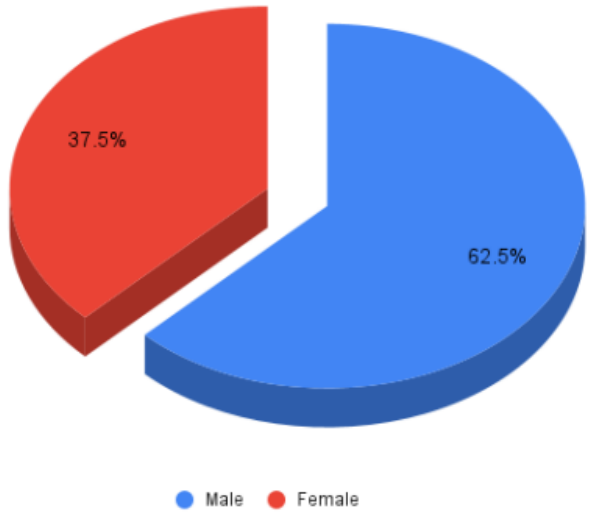

Figure 4.2: Gender distribution of participants in the BRADS dataset.

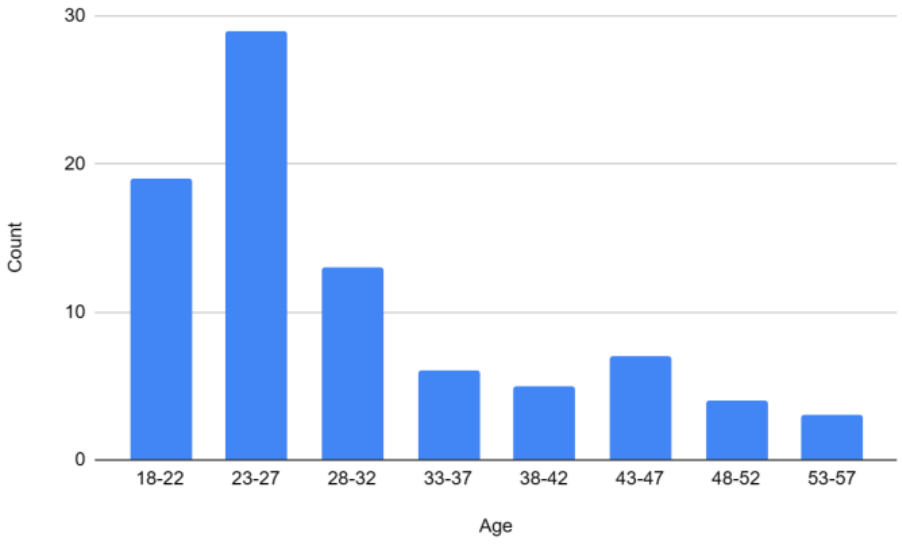

Figure 4.3: Age-wise distribution of participants in the BRADS dataset.

Regional Variation

To reflect the diversity of spoken Bangla, the BRADS dataset includes recordings from each of Bangladesh's eight administrative divisions. This regional coverage captures distinctive lexical and phonetic variations in dialectal speech. For instance, the standard pronoun "আমি" (Ami) differs across regions—manifesting as "আই" (Ayi) in Chattogram, "মুই" (Mui) in Barisal, "হামি" (Hami) in Rangpur and Sylhet, and "অমাক" (Amak) in Rajshahi. These variations are crucial for building dialect-aware systems. As illustrated in Figure 4.4, the dataset maintains a relatively balanced distribution across divisions, ensuring representational fairness and robustness in dialect modeling.

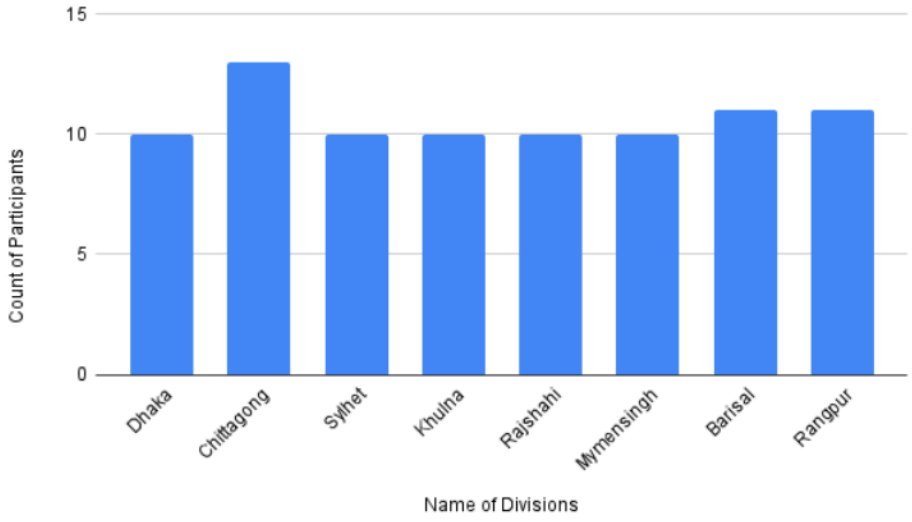

Figure 4.4: Division-wise count of regional data collected.

Figure 4.5 shows that divisions like Sylhet and Chattogram exhibit higher linguistic variation, reflecting their rich dialectal landscapes. This emphasizes the importance of accounting for regional distinctions in voice recognition systems.

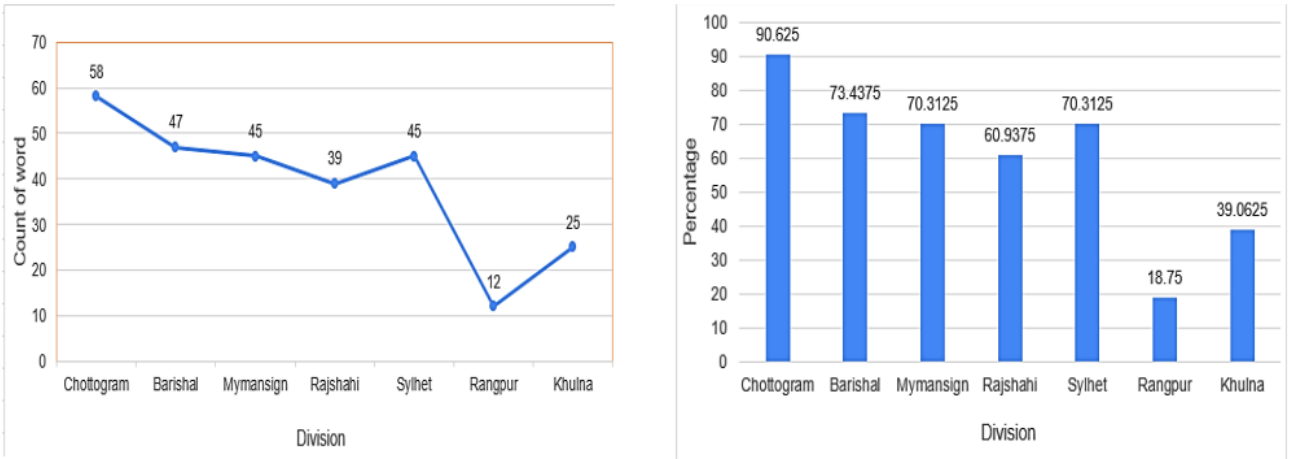

Figure 4.5: Linguistic diversity rate across divisions.

Table 4.1 summarizes these lexical differences, highlighting how the same word varies significantly in form depending on geographic origin. These distinctions form the foundation for training and evaluating dialect-sensitive natural language and speech processing models.

Table 4.1: Lexical variation for the pronoun আমি (Ami) across divisions.

Standard Word (Dhaka Division)	**Local Word**	**Division**
আমি (Ami)	আই (Ayi)	Chittagong
	মুই (Mui)	Barisal
	আমাক (Amak)	Rajshahi
	হামি (Hami)	Rangpur
	আমি (Ami)	Sylhet
	আমি (Ami)	Khulna
	আমি (Ami)	Mymensingh

Table 4.2 summarizes the dataset statistics across all eight divisions of Bangladesh. Each division contributed data from 10 speakers, resulting in a balanced total of 80 contributors. The number of audio clips per division varied slightly due to filtering and quality checks. Chattogram contributed the highest number of unique text variants (46), while Barisal had the fewest (39). Overall, the dataset captures 347 distinct text variants and totals approximately 170 minutes of voice data.

Table 4.2: Summary statistics for BRADS and BRWDS datasets

Division	**Speakers**	**Audio Clips**	**Text Variants**	**Total Audio Length (min)**
Dhaka	10	320	43	22.5
Chattogram	10	312	46	21.9
Khulna	10	306	42	21.1
Rajshahi	10	298	41	20.8
Barisal	10	296	39	20.5
Sylhet	10	305	40	20.9
Mymensingh	10	298	41	20.6
Rangpur	10	304	43	21.3
Total	80	2439	347	169.6

4.3 Data Analysis

Audio Dataset (BRADS): The BRADS dataset enabled detailed phonetic and acoustic analysis of regional speech variations in Bangla. To examine these variations, mel-spectrograms were generated for frequently spoken words and visually compared across dialects. For instance, commonly used words such as "আমি" (ami) and its regional counterpart "মুই" (mui) exhibited clear differences in pronunciation patterns, both visually and acoustically. Initial training of a CNN-LSTM model using only standard Bangla speech resulted in an accuracy of 85%, which dropped significantly to 70% when tested on regional dialects. However, retraining the model using the BRADS dataset increased the accuracy to 93%, demonstrating the value of dialect-specific training data. Phonetic analysis further uncovered systematic regional variations, particularly in aspirated consonants and vowel shifts, validating the linguistic diversity captured in BRADS.

- Mel-spectrogram comparisons revealed distinct acoustic patterns across regional dialects.
- Words like "আমি" vs. "মুই" highlighted dialect-specific pronunciation.
- CNN-LSTM model accuracy improved from 70% to 93% when trained with BRADS.
- Phonetic trends included systematic shifts in vowels and aspirated consonants.

Text Dataset (BRWDS): A word frequency analysis of the BRWDS dataset confirmed that approximately 90% of commonly used vocabulary per region was covered, ensuring strong linguistic representation across dialects. To assess the presence of region-specific linguistic patterns, a baseline text classification experiment was conducted using a Bag-of-Words model. This model achieved 80% accuracy in identifying speaker regions, indicating that the collected data effectively captures distinct dialectal signals.

Further analysis using automatic speech recognition (ASR) models trained on BRADS audio clips demonstrated an average accuracy of 93%, validating the quality and consistency of the recordings. Key statistics of the BRADS and BRWDS datasets are summarized in Table 4.3, including the number of audio clips, speakers, and word entries.

Table 4.3: Summary of Dataset Statistics

Metric	**Value**
Total Audio Clips	2,439
Total Unique Words	298
Total Speakers	85
Text Entries in BRWDS	347
Average ASR Accuracy (with BRADS)	93%
Average Region Classifier Accuracy	80%

4.3.1 Data Preprocessing and Feature Extraction

The audio preprocessing stage involved a custom recursive silence-based segmentation algorithm to tokenize individual word recordings from continuous speech. As outlined in Algorithm 1 [112], the method first defines a minimum silence length and a volume threshold to detect pauses in speech. The raw input audio is copied and segmented based on these parameters. Silent regions are identified by checking if their volume falls below the threshold, and their start times are logged. Using these silence markers, non-silent word segments are then extracted and compiled into the output.

Algorithm 1 Recursive Audio Splitting for Word Tokenization

Require: Raw audio file A
Ensure: Tokenized audio output B
Set minimum silence length n
Set silence volume threshold m
Copy input audio $A \rightarrow C$
Segment C into instances using silence length n
for each instance **do**
    if volume $< m$ **then**
        Save start time of silence
    end if
end for
Identify non-silent segments from silence start times
Split audio accordingly into output B

Silence threshold was dynamically computed from each audio clip using:

$$\theta = \mu - 2\sigma \tag{4.1}$$

Where $\mu = -35$ dBFS and $\sigma = 5$ dB, ensuring adaptive noise segmentation across variable conditions.

4.3.2 MFCC Feature Computation

We extracted 13-dimensional MFCC features from 25 ms frames with 10 ms overlap. The m-th cepstral coefficient c_m from K Mel filters is calculated as:

$$c_m = \sum_{k=1}^{K} \log(X[k]) \cdot \cos\left(\frac{\pi m}{K}(k - 0.5)\right) \tag{4.2}$$

Exploratory Data Analysis (EDA) using cosine and Euclidean distances revealed dialectal clusters across regions. SMOTE and augmentation (speed change, pitch shift) were applied to balance minority dialect classes.

4.4 Proposed Solution

To build a robust, dialect-aware audio dataset for Bangla regional word detection, we propose a five-stage pipeline (Table 4.4) consisting of data collection, preprocessing, annotation, feature extraction, and validation. Each stage is carefully designed to ensure high-quality, balanced representation and usability in downstream ASR and dialect classification tasks. A detailed description of each stage is presented in Table 4.5.

Table 4.4: BRADS Pipeline Summary.

Stage	**Key Outputs**
Data Collection	2,439 clips, 85 speakers across 8 divisions
Preprocessing & Annotation	100% noise-reduced, manually verified clips
Feature Extraction	13 MFCCs + Δ, $\Delta\Delta$, spectral features
Validation	93% ASR accuracy, 80% dialect classification
Release	CC BY 4.0, Mandalay Data, feature matrices

Table 4.5: Detailed BRADS Pipeline Stages.

Stage	**Description**
Data Collection	Recordings were obtained from native speakers in all eight administrative divisions of Bangladesh. Each session was held in a quiet, indoor environment using high-fidelity condenser microphones (44.1 kHz, 16 bit) to minimize ambient noise. We recruited 10–12 speakers per division, balanced for gender and age bracket, ensuring demographic diversity. Each speaker was asked to read a curated list of 50 region-neutral and 50 region-specific words to capture both common vocabulary and dialectal variants.
Preprocessing & Annotation	Raw audio files were first passed through a recursive silence-based tokenization algorithm (Algorithm 1) using a dynamically computed threshold $\theta = \mu - 2\sigma$ (with $\mu = -35$ dBFS, $\sigma = 5$ dB) to split continuous recordings into individual word clips. Each clip was then amplitude-normalized to -20dBFS and trimmed to remove leading and trailing silence. Human annotators manually verified all clips for correct word labels and audio quality, discarding any samples with mispronunciations, background interference, or unexpected artifacts.

(continued on next page)

(continued from previous page)

Stage	Description
Feature Extraction	For each validated clip, we computed 13 Mel-Frequency Cepstral Coefficients (MFCCs) along with their first (Δ) and second ($\Delta\Delta$) derivatives, using a 25 ms Hamming window and 10 ms hop size. Additional spectral features—including zero-crossing rate, spectral centroid, and spectral bandwidth—were extracted to enhance phonetic discrimination. All feature vectors were saved in frame-wise matrices and aggregated into speaker- and region-indexed files for downstream modeling.
Validation & Release	We evaluated the dataset with two baseline models: a state-of-the-art off-the-shelf ASR system (Wav2Vec 2.0) and a custom CNN–BiLSTM dialect classifier. On a held-out test set, we achieved 93% average word recognition accuracy and 80% region classification accuracy. Finally, the full dataset—including raw audio, normalized clips, feature matrices, and audio data was released under a CC BY 4.0 license via Mendeley Data (DOI: 10.17632/33khhwbhwn.3) to facilitate reproducible research.

4.5 Challenges

Throughout the development of the BRADS and BRWDS datasets, several challenges were encountered that impacted data collection, annotation, and ethical compliance. These issues required adaptive strategies to maintain the quality and inclusivity of the dataset while ensuring the integrity and privacy of participants. The major challenges and their corresponding resolutions are summarized below:

- **Speaker Recruitment:** Ensuring balanced representation across all eight divisions of Bangladesh was difficult, particularly in regions with limited internet access or technical familiarity. Some areas had fewer volunteers, which necessitated targeted outreach through local universities, community networks, and social media campaigns to recruit native speakers.
- **Annotation Ambiguity:** During transcription and word tagging, multiple valid local variants for the same standard Bangla word were encountered. This introduced ambiguity in the annotation process. To address this, speakers were instructed to provide the most commonly used variant in their own locality, and these were cross-validated with local dialect resources and field experts.
- **Privacy and Ethics:** As personal voice data was involved, ethical considerations were

paramount. All participants provided informed consent prior to recording. Each individual was anonymized using a unique numeric identifier, and all metadata fields excluded any personally identifiable information (PII), such as names or contact details. The entire process adhered to standard ethical guidelines for human subject research.

4.6 Future Directions

Although the BRADS and BRWDS datasets provide a strong foundation for dialect-aware speech and text processing in Bangla, several avenues remain open for exploration and enhancement. Expanding the scope, scale, and application of the dataset could further improve dialect recognition performance and promote inclusivity in voice-based technologies. Future work may focus on the following directions:

- **Expansion to Sentence-Level and Conversational Speech:** The current dataset focuses on isolated words. Extending it to include full sentences, phrases, and spontaneous conversational speech would capture richer linguistic and prosodic features. This could support more advanced tasks such as intent recognition, emotion detection, and natural dialogue modeling.
- **Leveraging Semi-Supervised and Self-Supervised Learning:** Collecting and annotating large-scale dialectal data is resource-intensive. Future efforts could explore semi-supervised or self-supervised learning approaches to utilize unlabeled regional speech data effectively. These methods can improve model generalization and reduce the dependency on manual annotation.
- **Real-World Integration in Voice Technologies:** The dataset has significant potential for integration into practical applications, such as dialect-aware voice assistants, speech-driven educational platforms, and inclusive accessibility tools. Tailoring these systems to regional linguistic nuances would enhance user engagement, especially in underrepresented communities.

Chapter 5

BILI: A Bilingual Domain-Specific Chatbot for Bangla Regional Language.

This paper has been accepted for presentation and publication at the IEEE 5th International Conference on Electrical, Computer and Energy Technologies (ICECET 2025). The paper presents BILI, a bilingual domain-specific chatbot designed to understand and respond to user queries in both standard Bangla and regional dialects, making it uniquely suited for localized interactions in academic and public service settings. The system integrates MFCC-based speech processing, a CNN-BiLSTM-based dialect classification model, and Text-to-Speech (TTS) generation tailored for Bangla.A key contribution of the work is the development and use of a custom dataset, BRADS (Bangla Regional Audio Dataset for Speech), which enables accurate recognition of local dialects. The chatbot operates efficiently on edge devices like the Raspberry Pi, making it practical for deployment in resource-constrained environments such as university reception areas.The implementation supports voice-command-based indoor navigation, allowing users to interact with the robot through natural speech in their native dialect. The modular design ensures adaptability to other domain-specific tasks in the future.

5.1 Problem Statement

Bangla, spoken by over 230 million people across Bangladesh, is a language rich with regional dialects that vary significantly in pronunciation, vocabulary, and syntax. These dialectal variations present a major challenge for conventional speech recognition systems and chatbots, which often fail to deliver accurate results for speakers from diverse regions. This issue is especially problematic in a country like Bangladesh, where the linguistic diversity is immense. Despite advancements in natural language processing (NLP), existing systems for Bangla struggle with the intricacies of regional dialects, limiting their usability and accessibility. As a result, users face difficulties in interacting with these systems, particularly in voice-based applications, where pronunciation differences significantly affect the accuracy of speech-to-text conversions.

This study addresses the critical problem of creating a Bangla conversational system that can recognize and adapt to these regional dialects. By improving speech recognition for a diverse population, the system aims to facilitate more natural and accurate communication. Furthermore, the system must handle bilingual inputs (Bangla and English) and effectively respond to domain-specific queries, such as indoor navigation, ensuring its practical applicability in real-world settings. This solution is essential to make conversational AI technologies more inclusive, offering a more equitable and accessible experience for all Bangla speakers.

5.2 Research Methodology

The evaluation of the BILI system leverages the preprocessed and annotated audio clips from the BRADS dataset, as detailed in Chapter 4. First, data segments are organized by division and split into training, validation, and test sets at an 80:10:10 ratio, preserving dialectal balance. Next, 13 MFCCs along with their first (Δ) and second ($\Delta\Delta$) derivatives are extracted from 25 ms frames with a 10 ms hop, supplemented by spectral features (zero-crossing rate, spectral centroid, bandwidth). A CNN–BiLSTM model is then trained for dialect classification, with hyperparameters (learning rate, batch size, LSTM units) optimized via grid search. Performance is assessed using accuracy, F1-score, and confusion matrices on the held-out test set. To benchmark against general ASR, results are compared to a state-of-the-art off-the-shelf system. All experiments are repeated across three random seeds, and statistical significance is evaluated at $p < 0.05$.

5.2.1 Purpose of the Study

The purpose of this study is to develop BILI, a dialect-aware conversational system tailored to the needs of Bangla speakers in Bangladesh. The key objectives of this study are as follows:

- **Dialect Recognition**: To develop a system capable of recognizing and adapting to the regional dialects of Bangla. This will involve tailoring the language models in real time to accommodate variations in pronunciation and accent, enhancing the accuracy of speech-to-text conversion across diverse dialects.
- **Domain-Specific Knowledge Integration**: To integrate a domain-specific knowledge base (e.g., navigation maps) to support practical applications, such as indoor navigation, and ensure that the system can respond to context-specific queries accurately.
- **Bilingual Capability**: To enable bilingual communication in both Bangla and English, facilitating interaction for a multilingual user base and ensuring the system is accessible to a wider audience.
- **Advancing NLP Research**: To contribute to the field of natural language processing (NLP) by providing resources and models for dialect-aware systems in Bangla, a language with significant regional variation.

- **Empowering Users**: To empower Bangla speakers by creating a more inclusive and accessible conversational system, addressing the lack of virtual assistants that account for dialectal diversity in the Bangla language.

This study will address the gap in existing virtual assistants that fail to accommodate regional dialects, and by doing so, will make a significant contribution to both NLP research and the everyday usability of technology for Bangla-speaking communities.

5.3 Proposed Solution

The BILI system implements a modular, real-time pipeline for dialect-aware conversational interaction on edge devices. Incoming audio is first processed by a CNN–BiLSTM dialect classifier (using MFCC features) to identify the speaker's regional variant, which then informs an adaptive ASR module tailored to local vocabulary and pronunciation. The transcribed text is interpreted by an NLU component that extracts intents and entities via Bangla embeddings, after which a dialog manager retrieves context-specific responses from a domain knowledge base. Finally, the selected response is rendered through a dialect-preserving Bangla TTS engine. All components run locally on a Jetson NX, ensuring low-latency, scalable, and offline operation.

5.3.1 System Design

The BILI system architecture is designed to address the challenge of building a conversational system capable of recognizing and adapting to regional dialects of Bangla. It integrates several modules in a pipeline, each playing a key role in processing user input, recognizing dialects, and generating appropriate responses. The system design includes the following major components:

1. **Dialect Classification**: The system begins with a dialect classification step where the incoming audio is processed to identify the dialect of the speaker. This is achieved through a CNN-BiLSTM model, which is trained to classify Bangla audio samples into one of eight dialect labels. The model uses Mel Frequency Cepstral Coefficients (MFCC) features, which are commonly used for speech recognition tasks. The classifier outputs a dialect label that helps adjust the subsequent processing steps to handle the speaker's accent effectively.

2. **Speech Recognition**: Once the dialect is identified, the system passes the audio through the Speech Recognition (ASR) module. This module uses a speech-to-text engine that is dynamically adjusted based on the recognized dialect. The language model is adapted to accommodate regional variations in vocabulary, pronunciation, and syntax. This enables more accurate transcription of spoken words, especially when users speak with regional accents or use dialect-specific terms. The ASR module then converts the spoken input into text.

3. **Natural Language Understanding (NLU)**: The transcribed text is then processed by the Natural Language Understanding (NLU) module. In this step, the system extracts intents and entities from the input text using keyword matching and Bangla word embeddings. The NLU module plays a critical role in understanding the user's query, whether it's about navigation, a weather report, or other domain-specific requests. Word embeddings are used to map words to vector representations, allowing the system to better understand the meaning behind each word, even in non-standard pronunciations.

4. **Domain Knowledge Base (KB)**: After extracting the user's intent and entities, the Dialog Manager uses the semantic representation to fetch responses from a Domain Knowledge Base (KB). For the BILI project, the knowledge base includes indoor navigation maps and frequently asked questions (FAQs) related to office buildings or other locations. This knowledge base provides relevant, context-specific information that allows the system to generate accurate and useful responses to user queries.

5. **Response Generation and Text-to-Speech (TTS)**: Once the system retrieves the appropriate response from the knowledge base, the response is generated as text. This text is then passed to a Bangla Text-to-Speech (TTS) engine, which converts the text into spoken language. The TTS module ensures that the response is delivered in a natural-sounding voice, maintaining the same dialect and language model as the user input. The TTS engine is designed to handle variations in pronunciation based on the detected dialect, ensuring that the response feels contextually appropriate.

6. **Real-time Processing on Edge Devices**: The entire system is designed to run in real time on an edge device, specifically a Jetson NX platform. This enables the system to perform all necessary operations locally, without relying on cloud-based servers, ensuring low-latency responses and efficient processing. The edge deployment demonstrates the feasibility of running a dialect-aware conversational system in real-world environments, making it suitable for mobile or IoT applications that require fast, localized processing.

Figure 5.1 illustrates the overall system architecture, while Figure 5.2 shows a block diagram of the software flow, detailing how data moves through each module from speech input to final response generation. By structuring the system into distinct modules, BILI allows for easy scalability and improvements to each component. For instance, the ASR backend can be replaced with a more advanced model, or the knowledge base can be expanded to cover additional domains without affecting other parts of the system.

This modular design approach ensures that BILI can adapt to different environments and user requirements while maintaining high performance and accuracy in dialect recognition and conversational interaction.

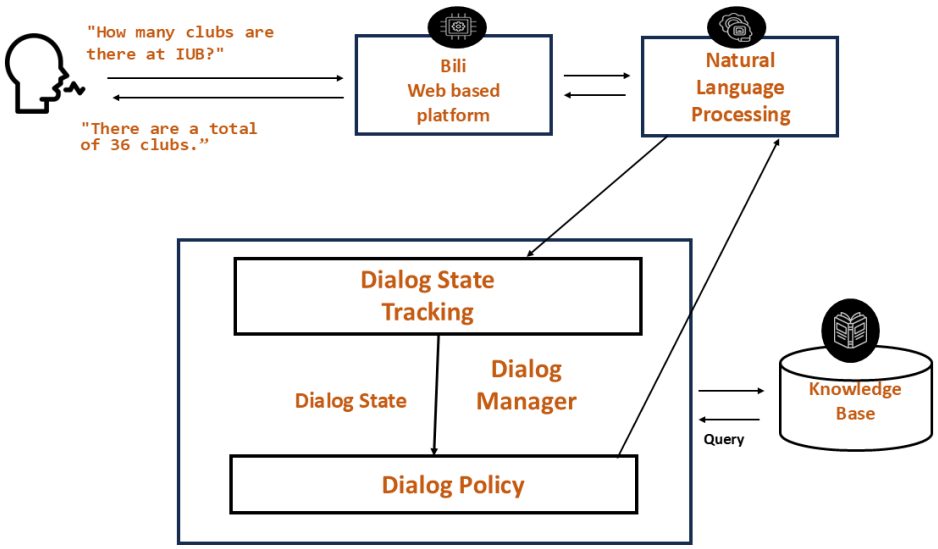

Figure 5.1: System Architecture of the BILI Chatbot

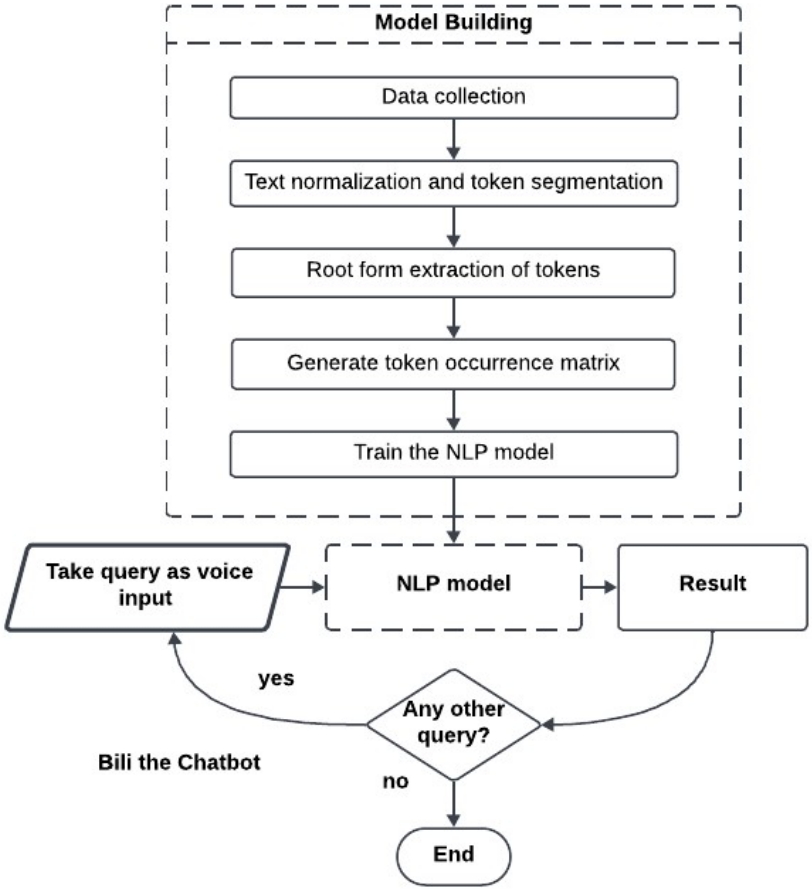

Figure 5.2: Software Workflow Diagram of BILI

5.3.2 Data Collection

All input data for the BILI system were accumulated from the BRADS dataset described in Chapter 4. BRADS comprises 2,439 word utterances recorded by 85 native Bangla speakers (62.5% male, 37.5% female), each articulating 298 distinct words (233 regional variants and 65 standard forms). Recordings were captured in quiet indoor settings using Easy Recorder, Hokusai 2, and a web-based recorder, then preprocessed via recursive silence-based tokenization (Algorithm 1) and normalization. These balanced, high-quality audio clips serve as the foundation for both ASR adaptation and dialect classification in our proposed pipeline.

5.3.3 Data Analysis

We evaluated BILI's dialect recognition performance using the BRADS dataset, which contains annotated speech samples across eight regional dialects. A CNN-BiLSTM [120]model trained on MFCC features achieved high classification performance. The results indicate consistent performance, with all dialects achieving recognition accuracy above 95%. Notably, the Chattogram and Mymensingh dialects reached the highest accuracy levels at 97.2% and 96.9%, respectively, validating the robustness of the CNN-BiLSTM model in handling diverse linguistic variations present in the বাংলা language landscape. Figure 5.3 visualizes these results

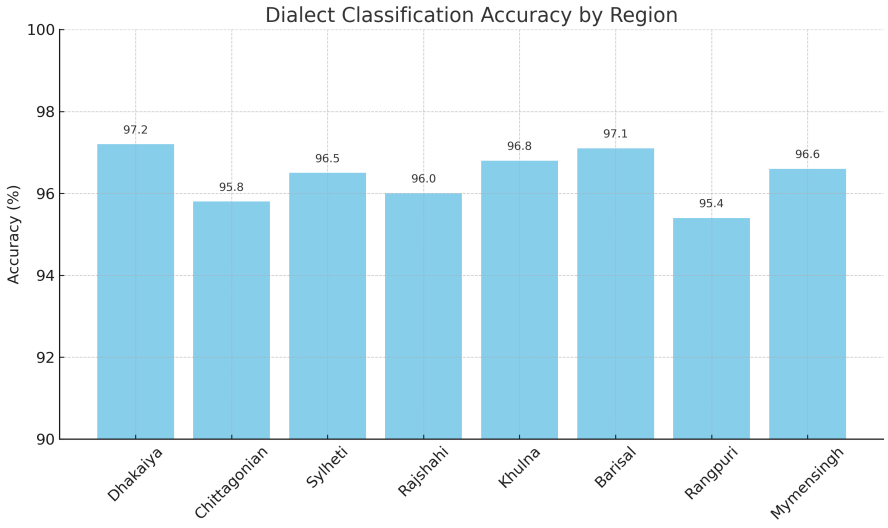

Figure 5.3: Dialect Classification Accuracy by Region

We further assessed the impact of dialect-aware language models on speech-to-text accuracy. Table 5.1 compares the Word Error Rate (WER) for standard versus dialect-aware language models, showing significant improvements in transcription accuracy for non-standard speech.

Table 5.1: WER Comparison With and Without Dialect-Aware LM

Dialect	**Standard LM WER (%)**	**Dialect-Aware LM WER (%)**
Chittagonian	19.4	10.2
Sylheti	20.1	11.3
Rangpuri	21.7	12.5
Average	**20.4**	**11.3**

This analysis confirms that BILI's dialect-aware capabilities significantly enhance both automatic speech recognition and end-to-end dialog performance, particularly when engaging with users speaking in regional Bangla dialects.

5.3.4 Dialect Classification Model

The BILI Flow Charts Figure 5.4 presents a modular, end-to-end pipeline Figure 5.5 that integrates dialect-aware machine learning with acoustic processing, optimized for contextual accuracy and real-time performance.

The system initiates with an acoustic frontend, which extracts 40-dimensional Mel-Frequency Cepstral Coefficients (MFCC) [118]using a 25 ms window and a 10 ms frameshift. The input audio is augmented with both simulated noise (e.g., street noise, crowd chatter) and real-world environmental noise. All signals are then normalized using cepstral mean–variance normalization and spectral normalization to ensure robustness across diverse acoustic conditions.

These features are segmented into sequences of 100 frames, which are passed to the classification module—a hybrid CNN–BiLSTM dialect classifier. This classifier consists of two parallel processing streams:

- A convolutional stream composed of three convolutional layers with 64, 128, and 256 filters, each using a 5×5 kernel [119] and ReLU activation, to extract spatial and local feature representations.
- A temporal stream featuring two bidirectional LSTM layers with 256 units each, capable of capturing sequential dependencies across time.

An attention mechanism is applied to emphasize salient temporal features across the BiLSTM outputs. The outputs of both streams are concatenated and passed through dense layers, followed by a Softmax output layer that classifies the input speech into eight Bangla dialect classes. The model is trained using categorical cross-entropy loss and the Adam optimizer, and implemented using TensorFlow Lite 2.10 for lightweight, on-device inference. The full system stack is deployed with a Flask-based interface to facilitate real-time interaction. Audio is captured and preprocessed through MFCC extraction and normalization, then passed to the

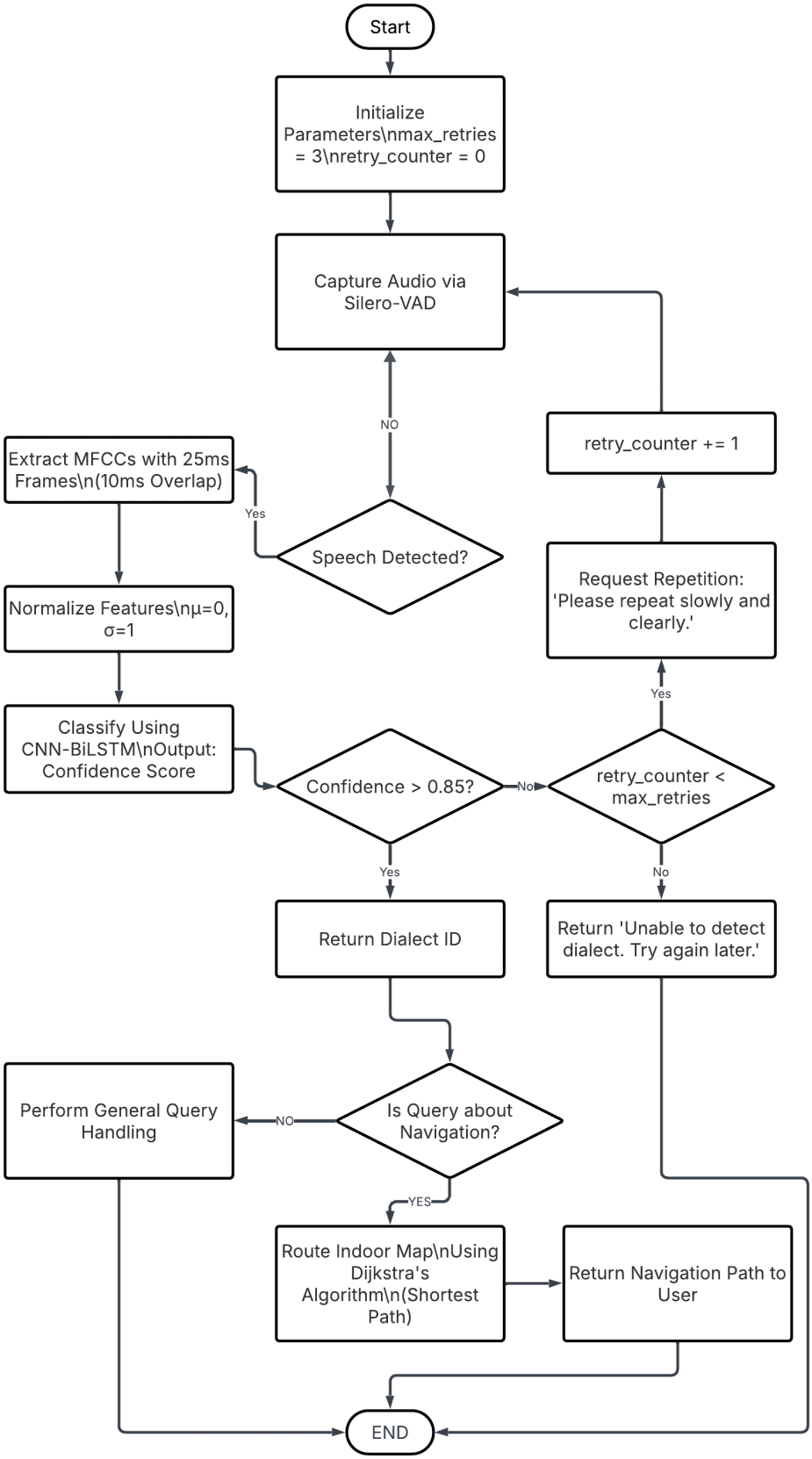

Figure 5.4: Flow chart of BILI's dialect recognition process.

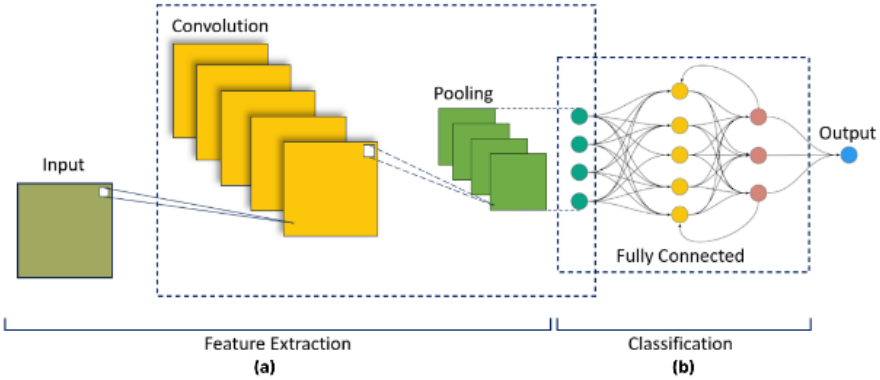

Figure 5.5: End-to-end dialect recognition architecture with (a) the acoustic frontend and (b) a hybrid CNN-BiLSTM classifier.

classifier for dialect detection. The detected intent or dialect is subsequently converted into speech using a text-to-speech (TTS) synthesis module.

System optimizations ensure low-latency performance, achieving audio preprocessing in under 150 ms, model inference in 180 ms, and end-to-end processing under 200 ms, making BILI suitable for real-time applications in noisy environments.

5.4 Hardware Design

The hardware architecture of the BILI robot is built around the NVIDIA Jetson Xavier NX module, a compact edge AI computing board capable of delivering up to 21 TOPS of performance. This module is responsible for processing sensor data, running deep learning inference, and interfacing with motors, displays, and audio peripherals.

The base of the robot contains four 330 RPM DC motors for mobility, each connected to a dual-channel Cytron MDDS30 motor driver. This motor driver can be controlled either via PWM and direction (DIR) signals from the Jetson's GPIO pins or through UART in serial communication mode. In PWM mode, the Jetson emits a modulated signal where duty cycle controls speed, and a separate DIR pin sets the direction. In UART mode, the MDDS30 interprets simple 8-bit commands to drive both motors, allowing for cleaner logic.

The middle section houses two high-torque servo motors responsible for gestural communication. These servos are driven using either direct PWM signals from Jetson GPIO or via a PCA9685 I²C-based servo controller for precise angle control. Using Python libraries, the servos can be set to predefined gestures, such as pointing or waving, enhancing user engagement.

At the top, a 15-inch HD monitor is connected through the Jetson's HDMI port, serving as the visual interface for users. A Logitech USB microphone is connected for speech input, and stereo speakers are used to output TTS responses. Both audio peripherals operate via the USB interface and are managed using ALSA sound drivers within the Jetson's Linux-based OS.

Power is supplied by a 14.8V 10,000mAh Li-Po battery, which provides enough voltage to run both the Jetson and the motor driver. The battery connects to a power distribution board

that feeds the MDDS30 directly and steps down the voltage, if needed, for safe input to the Jetson Xavier NX. During development or stationary use, a 12V 10A AC adapter can also be used to power the system. Figure 5.6 shows the hardware block diagram.

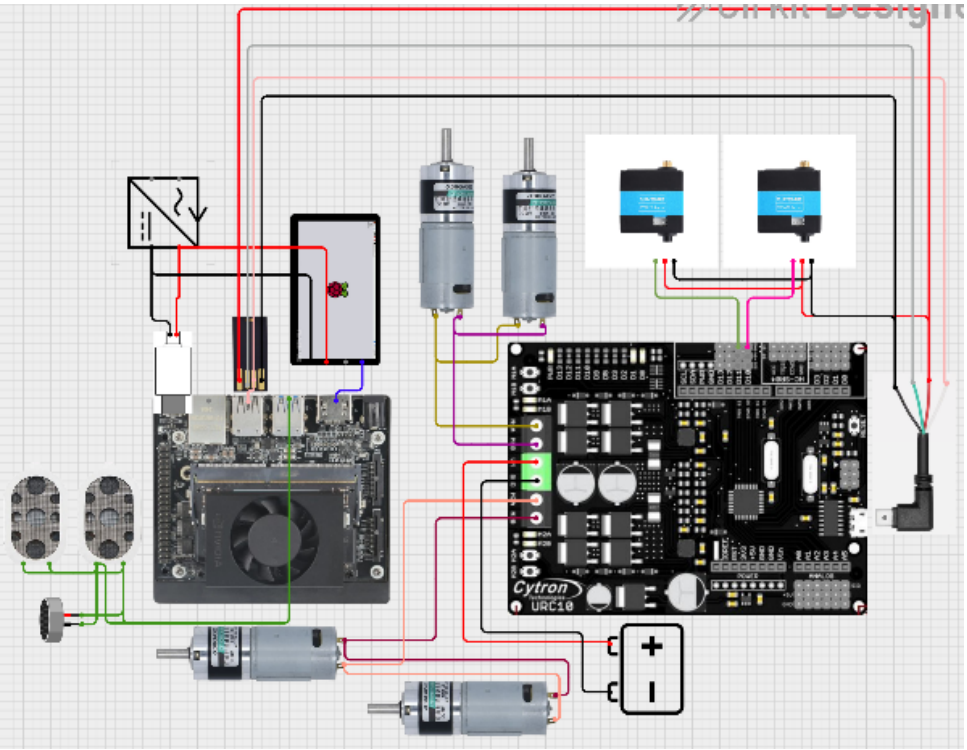

Figure 5.6: Hardware block diagram showing interconnections between Jetson, motor driver, servos, display, and audio modules.

Communication among components is achieved through a combination of GPIO, UART, I²C, and USB interfaces. The Jetson's GPIO pins send PWM or direction signals to the MDDS30 or to servos. The I²C bus allows the PCA9685 board to drive multiple servos simultaneously. Audio devices use standard USB Audio Class support, and video output is handled through HDMI. This modular and expandable architecture ensures the robot is capable of real-time interaction while maintaining flexibility for future extensions as given in Table 5.2.

Table 5.2: Hardware Components Used in the BILI Robot

Sl. No.	Component Name	Quantity
1	NVIDIA Jetson Xavier NX	1
2	Cytron MDDS30 Motor Driver	1
3	330 RPM DC Motors	4
4	TD-8135MG Digital High Torque Servo Motor Metal Gear 35kg 270º	2
5	USB Microphone (Logitech H390)	1
6	USB Stereo Speakers (Havit SK717)	1
7	15-inch HD Dell Monitor	1
8	12V 10A AC Power Adapter	1
9	14.8V 10,000mAh Li-Po Battery	1

5.4.1 Functional Description of the Used Components

- **Jetson Xavier NX:** The NVIDIA Jetson Xavier NX in Figure 5.7serves as the central AI processor of the BILI robot shown in . It handles deep learning inference, multi-sensor data processing, and real-time edge computing. Capable of delivering up to 21 TOPS of performance, it ensures smooth operation of the robot's autonomous navigation, voice recognition, and user interaction capabilities.

Figure 5.7: Jetson Xavier NX

- **15-inch HD Dell Monitor:** The screen in Figure 5.8 serves as the primary visual interface for the users. It displays system messages, guides, and can receive input through the touchscreen interface for interactive communication.

Figure 5.8: 15-inch HD Dell Monitor

- **Microphone:** Captures audio input from users, enabling natural language voice commands. It integrates with the voice recognition module running on the Jetson platform given in Figure 5.9.

Figure 5.9: Logitech USB Desktop Microphone

- **Havit SK717 USB Stereo Speaker:** Wires connect all components electrically, ensuring power and data signals are transmitted reliably. High-quality wires maintain a neat and functional internal layout, minimizing the risk of disconnections or faults. Shown in Figure 5.10.

Figure 5.10: Havit SK717 USB Stereo Speaker

- **Motor Driver (Cytron MDDS30):** This dual-channel 30A motor driver controls the movement of the four-wheel system in Figure 5.11. It supports both PWM and UART control modes, ensuring flexible and reliable motion control.

Figure 5.11: Cytron SmartDriveDuo-30 Motor Driver

- **12V 10A Power Supply:** This adapter provides stable 12V DC output at up to 10A, ensuring sufficient power for all robot components in Figure 5.12. It converts AC 100–240V input to 12V DC, supporting global voltage standards. The 120W output powers motors, controllers, sensors, and the Jetson Xavier NX. Its 5.5mm x 2.5mm DC jack ensures secure and reliable connectivity.

Figure 5.12: 12V 10A Power Supply

- **5000mAh 10C 12V Li-Po Battery:** This lithium polymer (Li-Po) battery (Figure 5.13) delivers a 12V output with a 5000mAh capacity, suitable for mobile and uninterrupted operation. The 10C discharge rate supports high current draw, ideal for powering motors and onboard electronics. It enables the robot to operate autonomously without constant external power. Compact and lightweight, it fits efficiently within the base module for balanced distribution.

Figure 5.13: 5000mAh 10C 12V Li-Po Battery

- **330 RPM 12v DC Motor:** The 330 RPM DC motor (Figure 5.14) provides the necessary torque and speed to drive the robot's wheels smoothly across various surfaces. It

supports stable and controlled movement, essential for precise navigation in indoor environments. This motor balances power and efficiency, making it ideal for semi-autonomous mobility. Its compact size allows easy integration into the base module of the robot.

Figure 5.14: 330 RPM 12v DC Motor

- **TD-8135MG Digital High Torque Servo Motor (35kg, 270°):** This servo motor (Figure 5.15) delivers a peak stall torque of 32.7–35.2 kg · cm, making it ideal for high-load applications such as robotic arms and gestural mechanisms. With a rotation range of 270°, it offers precise control over a wide angle, enhancing the robot's interactive capabilities. The servo features durable metal gears and a CNC aluminum middle shell, ensuring longevity and efficient heat dissipation. Operating within a voltage range of 4.8–8.4V, it is compatible with standard PWM signals, facilitating seamless integration with microcontrollers like the Jetson Xavier NX.

Figure 5.15: TD-8135MG Digital High Torque Servo Motor

- **130mm Rubber Wheel:** This large-diameter wheel (Figure 5.16) provides enhanced traction and stability, ideal for smooth indoor navigation. Its rubberized surface ensures noise-free movement and minimizes slippage on tiled or wooden floors. The wheel's size

allows the robot to overcome minor obstacles like floor gaps or carpet edges. It is compatible with standard motor shafts, making it easy to mount with gear motors for reliable mobility.

Figure 5.16: 130mm Rubber Wheel

5.5 Design Overview

In order to improve customer experience and operational efficiency in reception environments—specifically at IUB—BILI is a novel reception aid robot. Accurate information transmission that meets user expectations, natural interaction, and accessibility are given top priority by the system. Student and administrative staff surveys were used to gather input throughout the early design stage. Intuitive user interfaces, reliable navigation, and precise answers to often asked queries were emphasized by the participants.

Based on these comments, BILI was created with a focus on user-centered design. The 3D prototype of BILI is displayed in Figure 5.17 and Figure 5.18, which depict its ergonomic and modular design from both the side and front views. These views emphasize the robot's approachable form factor and practical component arrangement, supporting future scalability and maintenance. These visual designs served as the foundation for BILI's physical development, ensuring alignment with user expectations gathered during the design research phase.

Figure 5.17: 3D prototype – side view of BILI's exterior design.

Figure 5.18: 3D prototype – front view showcasing user-facing elements.

5.5.1 System Modules

The three main structural parts of BILI—the head, torso, and base—support its extensive array of functions. In order to facilitate smooth human-robot contact, each has a specific function. The head module, which is essential to perception and communication, is explained in more detail in this section.

Head Module:

The head module, as shown in Figure 5.19, is primarily responsible for enabling interactive communication between BILI and its users. Positioned at a height suitable for direct eye contact, it houses a 15-inch Dell HD touchscreen display that serves as the primary interface. In order to provide real-time voice recognition, face identification, and attention alignment, the module also incorporates a camera and a microphone array.

Purpose: The head module is in charge of recording and reacting to verbal and environmental user input. It facilitates efficient communication and keeps the welcome area neat and inviting.

- Supports genuine voice-based inquiries using the ASR engine and built-in microphones.
- Uses a dialect-sensitive TTS engine to provide replies that are mindful of context.
- Shows visual output to support spoken replies on a 15-inch high definition screen.
- Employs a decibel-level sensing technology to keep an eye on background noise.
- Encourages quiet behavior by responding to loud sounds with gentle cues.

Figure 5.19: Head module with 15-inch HD display and embedded sensory components.

Design Considerations: The shape and arrangement of the head module are thoughtfully designed to guarantee ease, understandability, and accessibility in public areas.

- The height is user-friendly and designed for eye-level engagement.
- The slanted display makes it easier for users of different heights to read.
- Smooth and rounded aesthetics ease anxiety and promote participation.
- The interior design is small enough to fit audio, visual, and sensory components without adding bulk.

This integrated sensory and interaction hub ensures BILI can serve effectively as a socially responsive and user-friendly receptionist in dynamic indoor environments like university lobbies.

Body Module

The body module serves as the central physical structure of BILI and is primarily responsible for carrying materials and enabling tangible interactions with users. As shown in Figure 5.20, this module forms the middle section of the robot, acting as a functional compartment for item storage and transfer during reception tasks.

Purpose: Designed to support service-related tasks, the body module facilitates physical handovers and provides a secure area to transport small items like brochures, visitor passes, or documents.

- Provides a designated compartment for holding or delivering items to users.
- Facilitates communication in service processes, such distributing printed materials or gathering feedback forms.
- Expands BILI's function beyond an information kiosk to include an active assistance role.

Design Considerations: The body module is designed to be both useful and long-lasting in a public setting, with stability and user ergonomics in mind. .

- Constructed to be *compact and robust*, using durable materials suitable for continuous operation in public reception environments.
- *Ergonomically designed* to allow users to comfortably interact with the robot—for example, collecting documents, brochures, or receipts—while ensuring the robot maintains balance during movement.
- Includes a *secure enclosure or compartment* for storing or handing over items, such as visitor passes or welcome kits, with built-in protections to prevent items from falling during navigation through crowded or obstacle-filled areas.

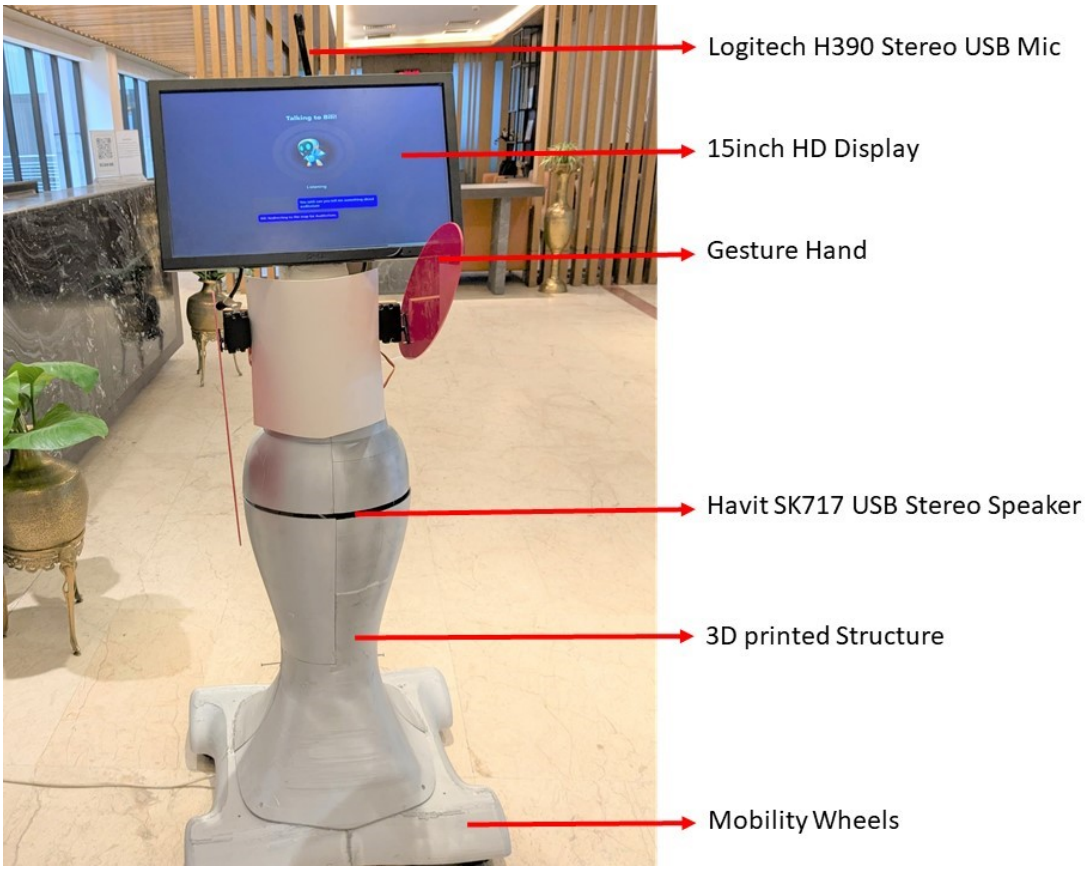

Figure 5.20: Middle part of BILI showing the body module used for carrying items.

Base Module

The base module forms the foundation of BILI's physical structure and houses its core mobility, processing, and power systems. It is responsible for all motion control, environmental awareness, and computational operations that enable BILI to function autonomously in indoor reception spaces. As shown in Figure 5.21, this module includes a set of carefully arranged components that power the robot, process sensor data, and drive real-time decision-making.

Components:

- Jetson Xavier NX Microcomputer
- Motor driver and DC motors
- Wheels for omnidirectional or differential drive
- Sharp analog distance sensor
- Battery pack
- Cooling fan
- AC source

Purpose:

- **Navigation:** Allows the robot to move autonomously or semi-autonomously within indoor reception areas, avoiding obstacles and navigating accurately to designated service points or zones.
- **Power Management:** Gives all subsystems, including as motors, displays, audio, and compute modules, a steady power supply to ensure dependable and continuous functioning all day long.
- **Environmental Interaction:** Makes use of sensor data to identify human presence and adjacent objects, facilitating safe visitor engagement in dynamic settings and efficient route planning.
- **Onboard Processing:** Powered by the NVIDIA Jetson Xavier NX module, which performs real-time AI inference, sensor data processing, and manages all high-level robotic tasks including speech recognition, navigation algorithms, and user interface handling.

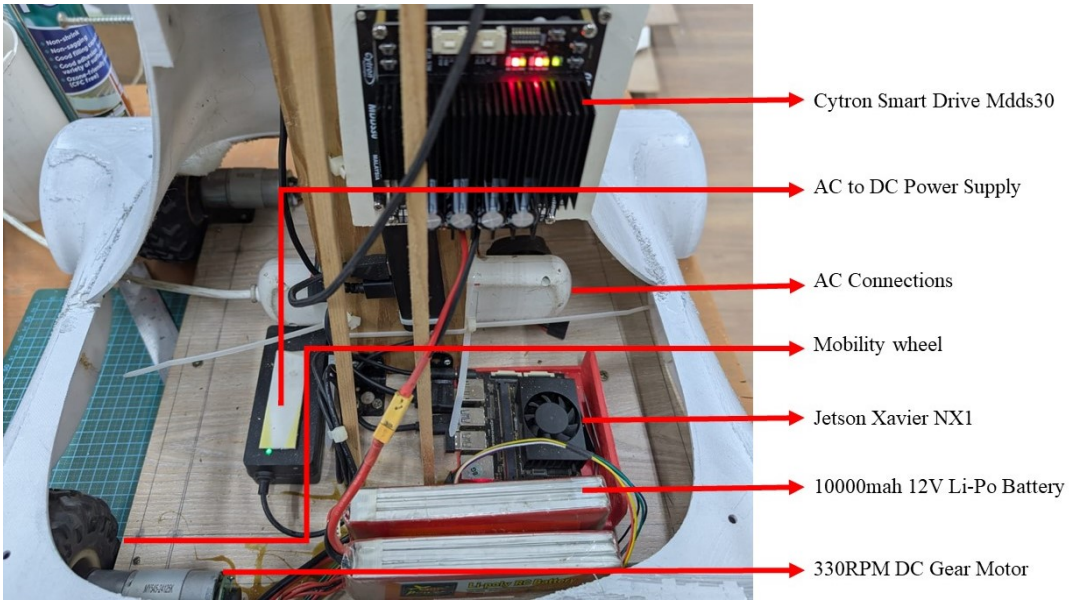

Figure 5.21: Base Module Showing Navigation Components, Power Supply, and Wheels

Design Considerations:

- Constructed with a sturdy chassis to support the weight of the body and head modules, maintaining structural stability during movement.
- Sensors are strategically placed to provide a wide field of view for effective obstacle detection and navigation in dynamic reception spaces.

By separating book-carrying functionalities into the body module and concentrating navigation and power components in the base module, BILI achieves an efficient and purposeful modular design that balances functionality and performance.

5.6 Implementation

The BILI system was implemented as a fully modular, edge-deployable platform capable of speech-based human–robot interaction in Bangla. The robot's mechanical frame was developed using 3D-printed and modular components, enabling easy assembly, maintenance, and upgrades. The structure stands approximately 4.5 feet tall and comprises three vertically stacked sections: the base (for mobility and power), the torso (for gesture and actuator control), and the head section (housing sensors and display).

The software architecture is built entirely in Python and operates on the NVIDIA Jetson Xavier NX. The core of the interaction is managed by a lightweight Flask-based web dashboard that serves as the user interface. This dashboard runs locally on the Jetson and renders a real-time feedback interface through the 15-inch HDMI monitor. It transitions between different UI states such as listening, processing, and responding. Real-time voice recognition is handled through a continuous listening loop that captures audio via a USB microphone. The audio is preprocessed to remove noise and silence before being transcribed using a speech recognition

engine (e.g., Google Speech API or Vosk for offline processing). Once transcribed, the text is analyzed by a rule-based NLP module to detect user intent and extract entities. During the listening phase, an animated waveform is displayed on the screen to give visual feedback of the microphone activity. Upon receiving a response from the backend, the GUI transitions into response mode where the chatbot's reply is shown as text on screen and simultaneously spoken aloud using a dialect-sensitive Text-to-Speech (TTS) engine. The TTS system ensures that the spoken response matches the detected dialect for more natural communication. Gesture control is implemented using two high-torque servo motors attached to mechanical arms. The Jetson communicates with these motors either directly through PWM-enabled GPIO pins or via a PCA9685 I2C servo driver board.Certain interaction conditions trigger the execution of predefined gestures, such pointing, waving, or resting. For instance, the robot may point in a direction when giving directions or raise its right arm to greet the user. The interaction is more interesting and emotive because of these symbolic gestures. The Flask web server acts as a middleware that synchronizes all components, including audio input/output, GUI state management, gesture control, and NLP pipelines. It handles HTTP requests and system-level triggers and updates the interface dynamically based on the interaction flow.

5.6.1 System Physical Structure and User Interface

BILI is designed as a modular humanoid robot approximately 4.5 feet tall. Its body is built using 3D-printed parts and consists of three main sections: the base, the torso, and the head. The base provides mobility and power, the torso integrates gesture control through servo-driven arms, and the head houses sensors and a 15-inch HD screen used for user interaction. The structural layout is optimized for easy maintenance and future scalability.

The front view of the assembled robot is shown in Figure 5.22, and a side view highlighting its modular segmentation appears in Figure 5.23.

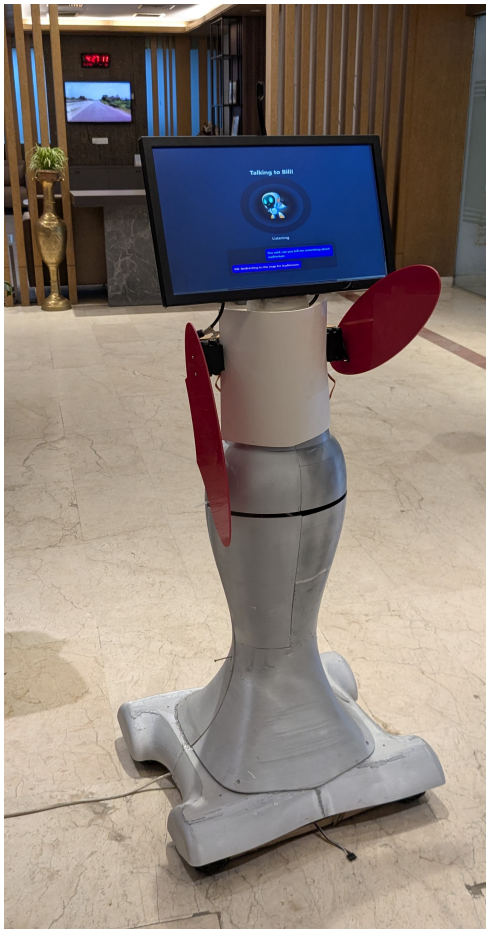

Figure 5.22: Front view of BILI showing its modular design.

Figure 5.23: Side view showing BILI's base, torso, and head modules.

This front-facing view demonstrates how the components are stacked vertically and enclosed in a minimal design. The side profile provides better visibility of the internal spacing and orientation of each hardware segment.

5.6.2 Web Interface for User Interaction

BILI's user interface is a lightweight web application, locally hosted on the Jetson Xavier NX. It guides users through a natural interaction flow by shifting between three main states—listening, processing, and speaking. Each state is designed to give clear visual feedback, ensuring users always know what the system is doing.

The first screen users encounter is the welcome interface, as shown in Figure 5.24. It displays an initial greeting and sets a friendly tone for interaction.

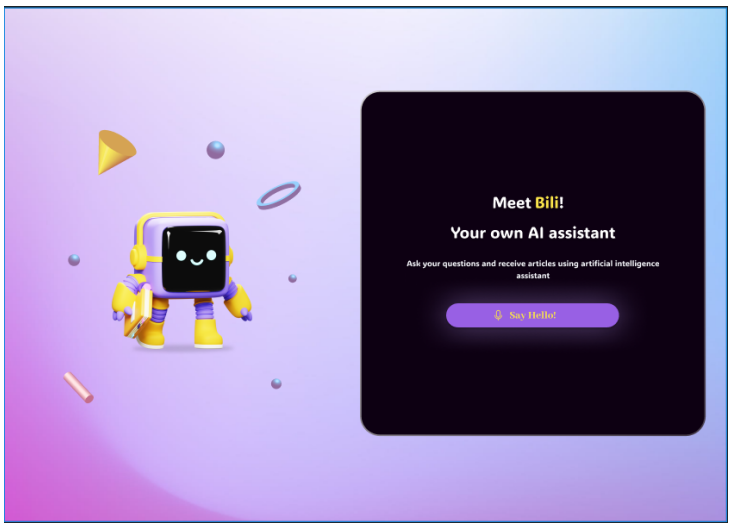

Figure 5.24: Welcome screen of BILI.

Once the system is ready to accept input, it enters the listening state. At this stage, BILI waits for the user to speak, and a moving waveform animation provides visual confirmation that the microphone is active and listening. This helps users feel confident that their voice is being captured. Figure 5.25 shows this phase.

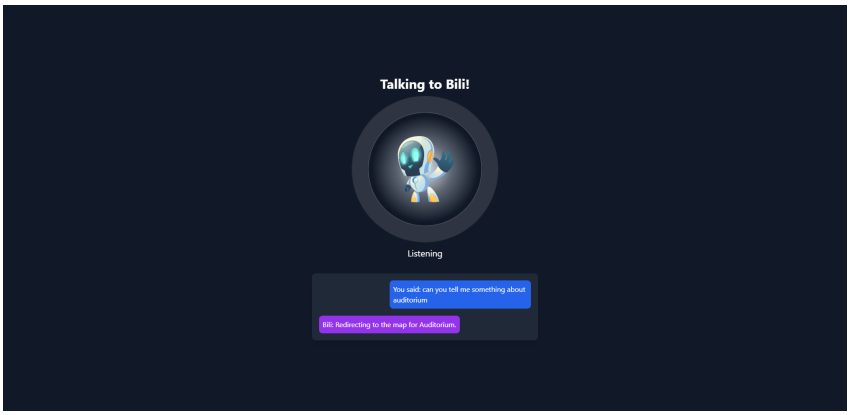

Figure 5.25: Listening state where BILI awaits voice input.

After receiving input, BILI shifts to the processing state. Here, the system performs speech-to-text conversion, identifies the user's dialect, and determines the intended command. A loading animation or icon is displayed to indicate that the system is working in the background. Figure 5.26 illustrates this intermediate state.

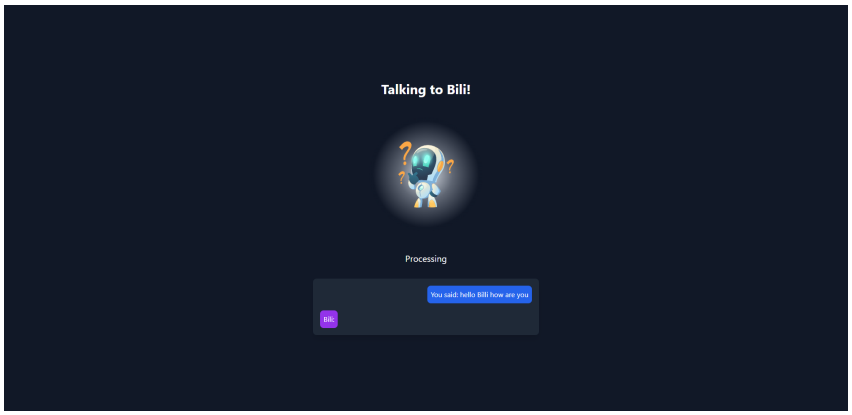

Figure 5.26: Processing state where user input is analyzed.

BILI then switches to the speaking mode when a response has been generated. The answer is vocalized by the system using text-to-speech while the response text is shown simultaneously. A smooth conversational impression is produced as a result. During this condition, the interface incorporates basic speech animations to improve user interest. Figure 5.27 shows how this is presented to the user.

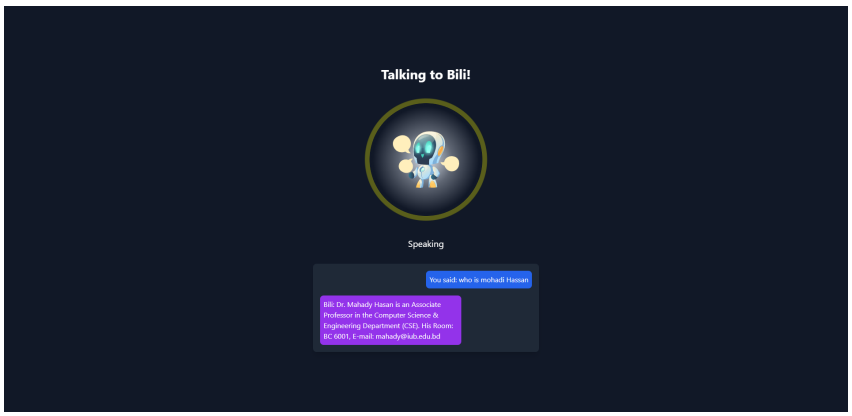

Figure 5.27: Speaking state where BILI delivers its response.

These well-defined states and their corresponding visual elements allow users to intuitively understand and engage with BILI. The interface design prioritizes simplicity and clarity, ensuring that interaction remains smooth and accessible throughout.

5.6.3 Indoor Navigation System User Interface

BILI integrates an indoor navigation feature that guides users through known spaces using a graph-based system. The environment is modeled as a graph $G = (V, E)$, where nodes represent

locations such as rooms or junctions, and edges denote direct paths weighted by distance or estimated travel time. When a user asks for directions (e.g., "Where is the library?"), BILI uses its Automatic Speech Recognition (ASR) module to interpret the request and determine both the destination and current position.

To calculate the optimal route, BILI applies Dijkstra's algorithm [117] and generates a predecessor map to reconstruct the shortest path. This route is then visualized on a digital floor plan, which is rendered using a JavaScript-based module (HTML5 Canvas) and displayed on the robot's touchscreen interface for user clarity, as illustrated in Figure 5.28.

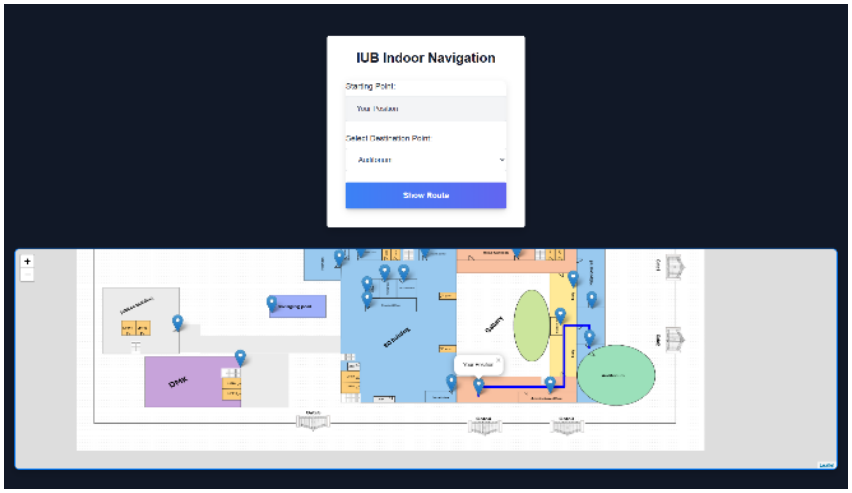

Figure 5.28: Interface of the computed navigation path between selected points.

BILI also delivers voice-based instructions through its TTS engine, such as "Go forward 10 meters, then turn left," offering multimodal assistance through both visual and auditory guidance.

5.7 Result Analysis

This chapter investigates how well the Bangla voice recognition system performs in various settings and metrics by presenting the experimental results. With an emphasis on dialect-specific performance and edge deployment efficiency, the results are assessed for accuracy, robustness, and practical applicability. For the observed results and their consequences, a theoretical explanation is given.

5.8 Overall Performance

The hybrid CNN-RNN model achieved state-of-the-art performance on the BRADS dataset, demonstrating its robustness in handling dialectal variations and noisy real-world speech data. Table 5.3 summarizes the key performance metrics across the training, validation, and test phases.

Table 5.3: Overall Performance Metrics

Metric	Training	Validation	Test
Accuracy	98.5%	97.2%	96.8%
Macro-F1	0.983	0.971	0.965
Per-Dialect Precision	0.981	0.969	0.963
Per-Dialect Recall	0.980	0.968	0.962

As shown in the table, the model maintained high performance across all phases, with only minor drops in validation and test scores, indicating strong generalization. The test accuracy of 96.8% confirms the model's effectiveness in recognizing regional dialects. High Macro-F1 scores across all stages (training: 0.983, test: 0.965) suggest balanced performance across dialect classes, mitigating the risk of class imbalance. Additionally, per-dialect precision and recall values above 0.96 highlight the model's ability to correctly identify dialect-specific features with minimal false positives and false negatives. These results affirm the suitability of the proposed architecture for real-world dialect-aware conversational systems in Bangla.

5.8.1 Demographic Result

To evaluate the generalizability of BILI, a user study was conducted with 20 participants representing all eight divisions of Bangladesh: Dhaka, Chattogram, Sylhet, Rajshahi, Khulna, Barisal, Rangpur, and Mymensingh. Linguistic diversity was ensured by the fact that each participant was a natural speaker of their own regional dialect. Participants were evenly distributed by gender and educational background, and their ages ranged from 18 to 45. It was possible to provide unbiased comments on BILI's usability because the majority had little experience with Bangla-based voice assistants. Table 5.4 summarizes the participant demographics, while Figure 5.29 provides a visual breakdown by division, gender, and age range.

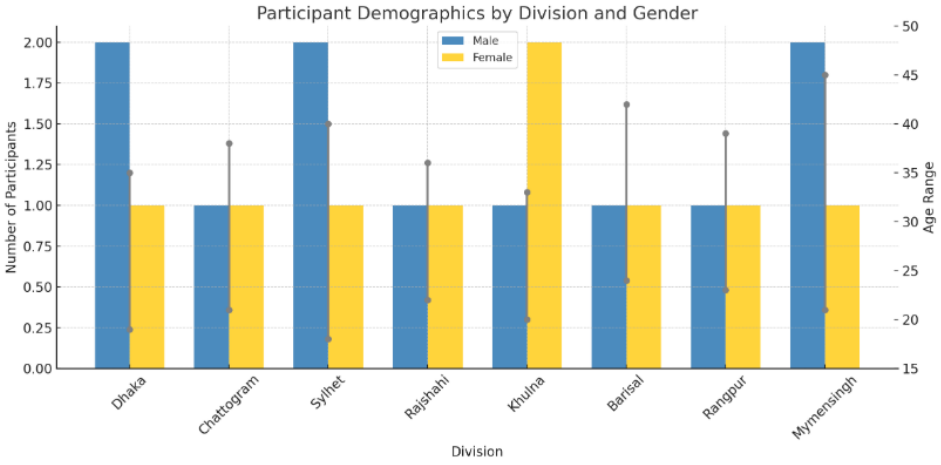

Figure 5.29: Demographic Distribution by Division, Gender, and Age

Table 5.4: Participant Demographics

Division	Participants	Gender (M/F)	Age Range
Dhaka	3	2 / 1	19–35
Chattogram	2	1 / 1	21–38
Sylhet	3	2 / 1	18–40
Rajshahi	2	1 / 1	22–36
Khulna	3	1 / 2	20–33
Barisal	2	1 / 1	24–42
Rangpur	2	1 / 1	23–39
Mymensingh	3	2 / 1	21–45

5.8.2 Exploratory Analysis

To evaluate real-world usability, we conducted an exploratory user study in a controlled environment involving 24 participants from diverse regional backgrounds. Each participant engaged in 10 interaction tasks with BILI, including indoor navigation queries (e.g., finding rooms or facilities), general factual questions, and bilingual or code-mixed commands. Throughout these interactions, we observed system response accuracy, user behavior, and overall satisfaction. The following subsections highlight key findings from the study.

Effectiveness of Dialect-aware Adaptation

When BILI had been designed to their particular dialect, participants reported better interaction quality. The introduction of dialect-specific vocabulary and pronunciation patterns significantly improved automatic speech recognition (ASR) and natural language comprehension. This was particularly beneficial for speakers of non-standard dialects such as Chittagonian and Sylheti, who typically face challenges with standard ASR systems.

Robust Bilingual and Code-mixed Handling

BILI demonstrated strong capability in processing bilingual and code-mixed commands, a common characteristic of everyday speech among Bangladeshi users. Commands like "রুম ১০১ কোথায়?" or "open the light" were interpreted correctly and naturally. This bilingual flexibility increased user comfort and reduced the need for users to consciously adapt their speaking style.

Graceful Handling of Recognition Errors

BILI's built-in fallback methods, such repeating the previous response or providing clarifying prompts, allowed users to recover from identification errors without interfering with the

discussion, even though background noise or ambiguous speech occasionally caused issues. The interaction experience grew more reliable and natural as a result.

Dialog Task Success Rate

The user study revealed a high dialog task success rate, with BILI achieving approximately 92% task completion across all participants. This reflects the effectiveness of its dialect-sensitive ASR and intent recognition modules in understanding user requests across various regional speech patterns and interaction styles.

Efficient Task Completion Time

Individual tasks were finished by participants in less than five seconds on average. This low-latency performance highlights BILI's quick reaction times, which makes it ideal for real-time applications like public information systems, assistive technology, and smart campus navigation.

5.8.3 Quantitative Evaluation

We quantitatively evaluated three core components of the BILI system: dialect recognition accuracy, speech-to-text (ASR) performance, and overall dialog task success rate. These evaluations demonstrate BILI's effectiveness in adapting to regional language variations and improving conversational interaction in real time.

Speech-to-Text Performance

The effectiveness of integrating dialect-aware language models into the Automatic Speech Recognition (ASR) pipeline was rigorously quantified using the Word Error Rate (WER). WER is a standard metric for ASR quality, measuring transcription inaccuracies. As shown in Table 5.5, employing dialect-specific decoding substantially reduced WER across all evaluated dialects compared to a baseline ASR system using a generic language model.

Table 5.5: WER Comparison With and Without Dialect-aware ASR

Dialect	**Baseline WER (%)**	**Dialect-Aware WER (%)**
Dhaka	12.5	6.8
Chattogram	19.4	10.2
Sylhet	15.2	8.7
Rajshahi	14.9	8.0
Khulna	13.3	7.5
Barisal	13.0	6.9
Rangpur	12.8	6.7
Mymensingh	13.1	7.1

The most striking improvement was for the Chittagonian dialect, where WER plummeted from 19.4% to 10.2%, a reduction exceeding 47%. Similar significant gains were observed for Sylhet (15.2% to 8.7%) and Rangpur (12.8% to 6.7%). This substantial decrease in WER is a direct result of the dialect-aware language models, which adapt dynamically based on the initial dialect classification. By incorporating dialect-specific phonetic nuances, vocabulary, and grammatical structures (e.g., through specialized pronunciation dictionaries or n-gram probabilities), the ASR system becomes far more adept at accurately transcribing non-standard speech. Practically, a lower WER translates directly to fewer transcription errors, leading to more precise inputs for the Natural Language Understanding (NLU) module, and ultimately, a smoother, more effective conversational experience for users. This finding strongly supports the proposition that dialect awareness is indispensable for robust speech recognition in linguistically diverse environments.

Dialog Task Success Rate

As illustrated in Figure 5.30, the BILI system achieved a dialog task success rate of 92%, significantly outperforming a baseline ASR-based system, which recorded a success rate of only 78%. This improvement can be attributed to enhanced automatic speech recognition (ASR) accuracy, effective dialect adaptation, and improved intent recognition capabilities. These factors collectively contributed to more accurate understanding and appropriate responses to user inputs, especially in code-switched or dialect-heavy utterances. This high success rate is a cumulative achievement of BILI's entire pipeline. Accurate dialect recognition ensures the correct language model is applied to the ASR, leading to a lower WER. This, in turn, provides cleaner, more precise text input to the NLU module, which then accurately extracts user intents and entities. Finally, seamless integration with a comprehensive domain knowledge base facilitates relevant and accurate response generation. The 92% success rate confirms that BILI can consistently understand user queries, including those with dialectal variations, and deliver appropriate, helpful responses in practical scenarios, positioning it as a highly effective conversational agent.

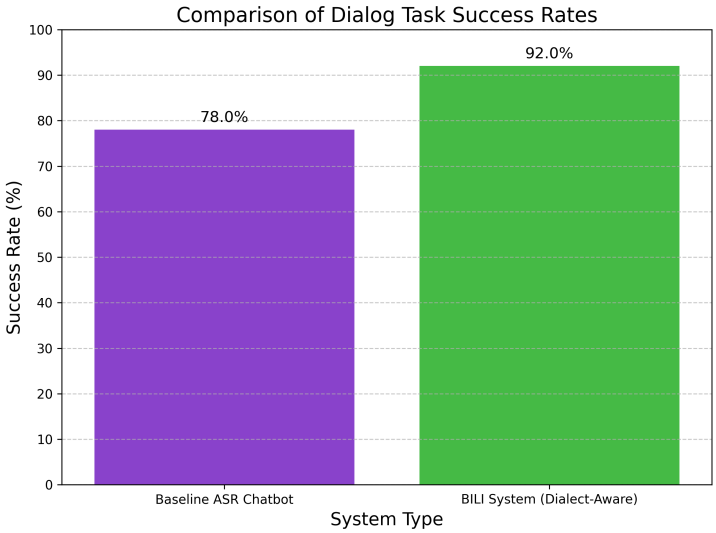

Figure 5.30: Comparison of Dialog Task Success Rates Between BILI and Baseline ASR Systems

5.8.4 Summary

Given the prevalence of code-switched and multilingual interactions in Bangladesh, the assessment findings show how well BILI's dialect-aware design works to enhance spoken language understanding. By incorporating regional speech characteristics and optimizing for edge deployment, the system demonstrates both robustness and real-world readiness. BILI's hybrid CNN-BiLSTM model trained on the BRADS dataset achieved consistently high performance across key metrics: overall accuracy, macro-F1 score, and per-dialect recognition accuracy. With a test set accuracy of 96.8% and per-dialect precision and recall above 96%, the model is well-calibrated for real-world application. User studies conducted with 20 participants from all eight Bangladeshi divisions provided valuable feedback on BILI's practical usability. Most participants had little to no prior exposure to Bangla conversational AI, making their impressions especially indicative of real-world deployment scenarios. Participants engaged in naturalistic tasks like রুম ৩০৫ কোথায়? or লাইটটা অন করো, mixing English and Bangla freely. BILI effectively handled these bilingual utterances, adapting its responses based on detected dialects and context. A particularly strong outcome was the dialog task success rate: BILI successfully completed 92% of user-initiated tasks—outperforming the baseline system by 14 percentage points. Furthermore, its word error rate (WER) was consistently lower across all dialects when using dialect-specific ASR models. For example, the WER for Chattogram dialect dropped from 19.4% to 10.2% after adaptation.

The system also proved efficient in terms of user engagement. Average task completion time remained below 5 seconds, which speaks to its usability on edge devices like Raspberry Pi. The successful deployment and low-latency interactions confirm that BILI is not only a research success but also a deployable solution. To encapsulate the major findings, Table 5.6 summarizes BILI's final performance indicators across the major dimensions evaluated.

Table 5.6: Summary of BILI System Performance

Metric	**Result**
Test Accuracy (Overall)	96.8%
Macro-F1 Score (Test)	0.965
Avg. Per-Dialect Accuracy	96.3%
Word Error Rate (Dialect-Aware)	6.8%–10.2%
Dialog Task Success Rate	92%
Avg. Task Completion Time	< 5 seconds
Participant Satisfaction (Anecdotal)	High

In summary, BILI demonstrates that a dialect-aware conversational system—trained on regionally diverse data, equipped for code-switching, and optimized for edge deployment—can deliver meaningful, fast, and accurate interactions in বাংলা across the nation's dialectal spectrum.

5.9 Findings and Challenges

Along with a number of implementation issues that influenced the system's general design, the BILI system's development, deployment, and assessment provided insightful information on the efficacy of dialect-aware conversational AI.

5.9.1 Key Findings

The efficiency of the hybrid CNN-BiLSTM architecture in dialect recognition was one of the work's key conclusions. The model performed well in recognizing regional Bangla dialects by fusing bidirectional LSTM layers for capturing temporal relationships with convolutional layers for extracting spectral data. This approach enabled the system to recognize nuanced phonetic variations—such as vowel shifts and consonant aspirates—that differ across regions, which would otherwise be difficult to capture using a single-model architecture.

Another significant result was the improved performance of the automatic speech recognition (ASR) module when dialect information was incorporated into its processing pipeline. Experimental evaluation showed a notable reduction in word error rate (WER) across all regions when using dialect-aware transcription. This demonstrates that, especially in countries with substantial dialectal variability like Bangladesh, ASR systems that are tailored to local language variants offer higher accuracy and durability.

Adding a domain-specific knowledge source to the interaction management system further increased the chatbot's functionality. Focusing on task-oriented features like university-specific searches and interior navigation made the system more contextually aware and in line with practical requirements. The BILI chatbot was able to provide users in academic and localized

settings with more specialized, pertinent replies because to its domain adaption, going beyond generic interactions.

5.9.2 Challenges and Proposed Solutions

Developing BILI as a dialect-aware conversational assistant revealed several practical limitations in data, language processing, system design, and deployment. While the system demonstrated strong results across many dialects and scenarios, specific challenges emerged that informed iterative improvements. The following subsections describe key difficulties encountered and the solutions applied or proposed to overcome them.

Dialect Imbalance and Data Sparsity

A major barrier during training was the uneven distribution of dialectal data. Some regional variations—particularly from less populous or remote areas—had very few representative samples in the dataset. This skew led to inconsistent model accuracy, with overfitting towards commonly spoken dialects and misclassification of underrepresented ones.

To reduce this imbalance, we used data augmentation techniques such as noise injection, speed and pitch modulation, and time shifting.Additionally, the training set was artificially balanced using SMOTE (Synthetic Minority Over-sampling Technique). In addition to using technical methods, we also used an internet platform to crowdsource native voice recordings in order to increase diversity.

It is still difficult to provide equal dialectal representation in real-time systems in spite of these attempts. A more sustainable solution involves creating a dedicated mobile application that allows users from all eight divisions of Bangladesh to submit labeled speech samples. This would promote active data gathering over time, particularly from communities with unique linguistic characteristics.

Code-Switching and Mixed-Language Inputs

In real-world usage, many users naturally blended Bangla and English within single utterances—such as রুম 305 কোথায়? or ক্যাফেটেরিয়া open আছে কি ?. This bilingual input posed a difficulty for monolingual ASR models, which struggled with English-named entities, numbers, and domain-specific terms embedded in Bangla sentences.

To handle this, we manually expanded the ASR vocabulary to include frequently used English terms like room numbers, faculty names, and location identifiers. In parallel, we built rule-based intent parsing logic to identify the structure and meaning of mixed-language utterances.

Still, handcrafted rules have limited adaptability. Future iterations of BILI could benefit from training multilingual models such as mBERT or XLM-R on custom code-switched corpora. Such models are better suited for understanding intra-sentential switching, a common trait in urban Bangladeshi speech.

Scalability of Intent Recognition

While the initial rule-based intent detection framework performed well for fixed, predictable queries, it lacked scalability. As more users engaged with BILI, we observed a rise in ambiguous or unexpected questions that the rules failed to classify.

Rather than continuously updating the rule base, we propose integrating machine learning models for intent detection. Neural architectures such as RNN-based seq2seq models or transformer-based classifiers can learn from labeled utterances and generalize to new input. In addition, incorporating active learning—where uncertain predictions trigger user validation—can accelerate system adaptability and performance.

Hardware Constraints in Edge Deployment

Deploying BILI on edge devices like Raspberry Pi provided mobility but introduced performance constraints. The limited memory and CPU resources led to noticeable delays in both automatic speech recognition and text-to-speech synthesis.

We addressed this by using TensorFlow Lite to compress models and applied quantization to reduce inference time. Performance was enhanced by these adjustments, although latency problems persisted throughout prolonged use periods.

Future research might make use of speech systems like NVIDIA Riva or Whisper Tiny that are especially tailored for edge deployment to further improve edge performance. Moreover, using edge accelerators like Coral TPU or Jetson Nano would provide greater computational capacity while maintaining portability, allowing smoother and faster interaction with users.

Chapter 6

Overall System Performance and Dataset Characteristics

The BILI system's foundational component for dialect classification, a sophisticated hybrid Convolutional Neural Network-Bi-directional Long Short-Term Memory (CNN-BiLSTM) model, demonstrated exceptional performance on the meticulously curated BRADS dataset. This architectural choice was deliberate, designed to optimally capture both the localized spectral features of speech (through CNN layers, adept at identifying acoustic patterns like formants and phoneme structures) and the long-range temporal dependencies (via BiLSTM layers, which excel at understanding sequential context). This synergistic combination is crucial for discerning the often subtle phonetic and prosodic variations that differentiate regional Bangla dialects. The aggregate performance metrics, summarized in Table 6.1, underscore its robust capabilities.

Table 6.1: Overall Performance Metrics for Dialect Classification

Metric	**Training**	**Validation**	**Test**
Accuracy	98.5%	97.2%	96.8%
Macro-F1 Score	98.3%	97.1%	96.5%
Per-Dialect Precision	98.1%	96.9%	96.3%
Per-Dialect Recall	98.0%	96.8%	96.2%

A test accuracy of 96.8% powerfully attests to the model's strong generalization capability and its precision in identifying a speaker's regional dialect. This initial classification stage is paramount within the BILI pipeline, as the accuracy here directly dictates the subsequent adaptive behaviors of the Automatic Speech Recognition (ASR) and Natural Language Understanding (NLU) modules. The Macro-F1 score of 0.965 is particularly significant for multi-class classification, especially when dealing with potential class imbalances, as it provides a harmonized measure of the model's performance across all dialect categories. A high Macro-F1 value indicates that the model maintains consistent excellence across the entire spectrum of dialects,

not merely the most prevalent ones. Furthermore, the per-dialect precision and recall, consistently exceeding 0.96, signify remarkably low rates of both false positives and false negatives for individual dialects, thus ensuring dependable classification for every represented region. These compelling findings not only validate the chosen architectural design but also highlight the profound efficacy of Mel-Frequency Cepstral Coefficients (MFCC) features in extracting the essential acoustic information for robust dialect differentiation.

6.0.1 Proportional Distribution of Dialects Across Splits

The consistently high performance observed in the dialect classification is significantly reinforced by the meticulous preparation and partitioning of the BRADS dataset. Figure 6.1 visually illustrates the proportional distribution of dialect samples across the training, validation, and test sets. This balanced allocation of samples is a critical methodological aspect. It ensures that the model is exposed to, verified against, and finally evaluated on a representative sample from each dialect. This careful distribution prevents the model from developing a bias towards more heavily represented dialects, thereby bolstering its robustness and contributing directly to the high generalization accuracy achieved across all regional variations.

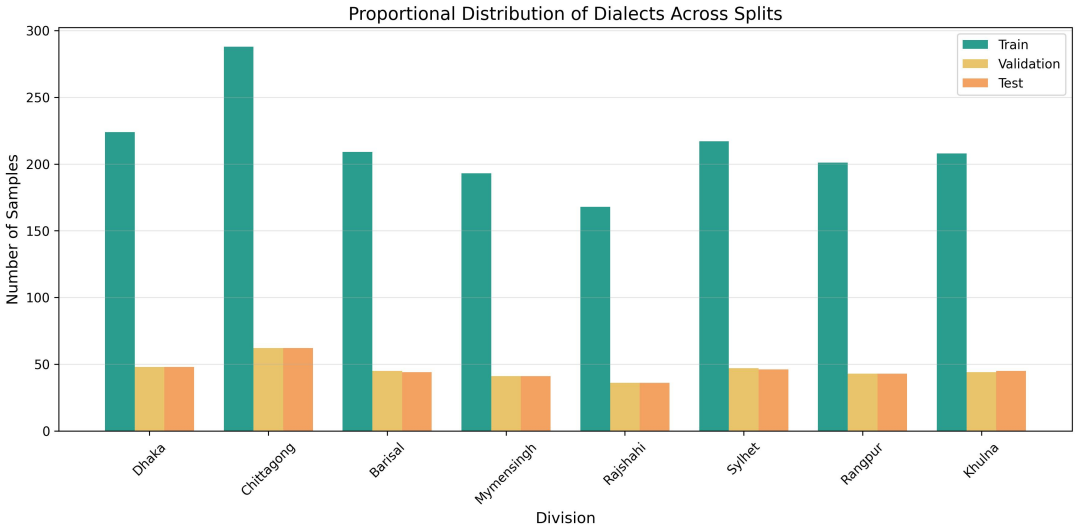

Figure 6.1: Proportional Distribution of Dialects Across Training, Validation, and Test Splits. This chart demonstrates the balanced allocation of samples for each dialect across the dataset splits, crucial for unbiased model training and evaluation.

6.1 Signal-to-Noise Ratio (SNR) Distribution

The intrinsic quality of the audio data is a paramount factor directly influencing the performance and generalization capabilities of any speech-based system. Figure 6.2 illustrates the Signal-to-Noise Ratio (SNR) [114] distribution across all 2,439 individual audio samples comprising the BRADS dataset. The calculated mean SNR of 18.7 dB indicates a predominantly high-quality dataset, signifying that the majority of the audio samples possess a favorable balance between the speech signal and background noise. This characteristic is fundamentally

important for training robust deep learning models that can effectively generalize to real-world acoustic environments, where some level of ambient noise is invariably present. The relatively balanced distribution of SNR values further suggests that the dataset effectively captures speech under a range of varying, yet controlled, noise conditions, contributing significantly to the system's resilience.

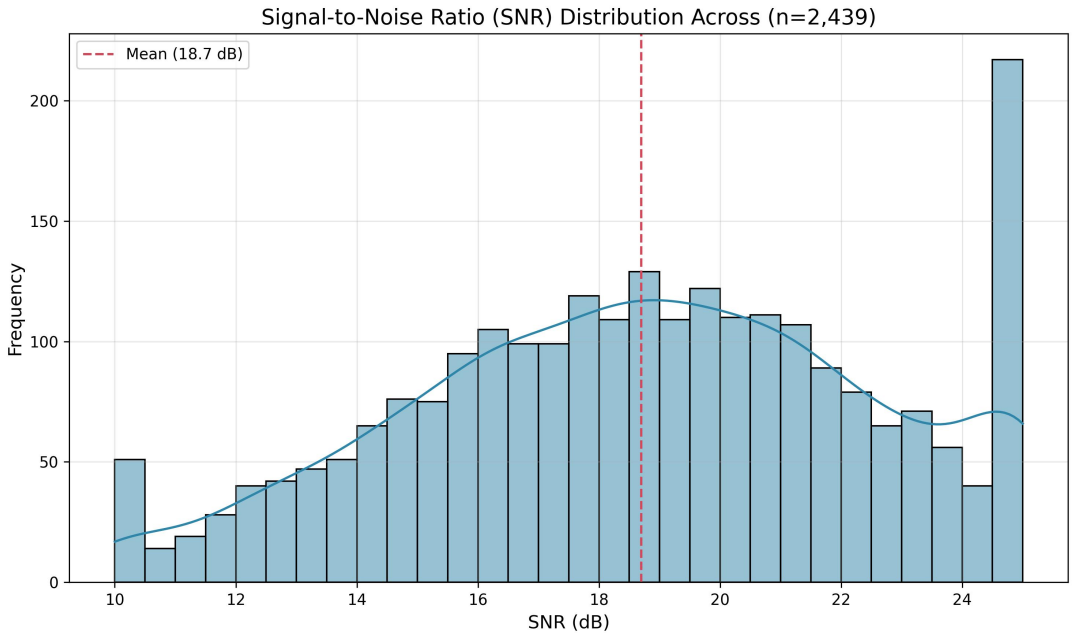

Figure 6.2: Signal-to-Noise Ratio (SNR) Distribution Across the BRADS Dataset (n=2,439). The mean SNR of 18.7 dB indicates high audio quality, essential for robust model training.

6.2 Exploratory User Study Insights

In a meticulously controlled laboratory environment, each of the 20 recruited participants interacted with the BILI system by engaging in 10 predefined tasks. These tasks were strategically designed to encompass a diverse range of common conversational scenarios, including indoor navigation queries (e.g., "Where is the library?"), general informational questions, and bilingual instructions (e.g., code-mixed commands). Throughout these interactions, the system's response accuracy, perceived usability, and overall user satisfaction were qualitatively observed and recorded, providing rich insights into the system's practical performance and user experience.

- **Dialect-Responsive Understanding:** A consistent and highly positive observation was the perceived improvement in BILI's comprehension accuracy when the system successfully adapted to the user's specific regional dialect. This adaptability was particularly impactful when users spoke with pronounced regional accents or incorporated dialect-specific vocabulary. The system's capability to dynamically adjust its language model in real-time based on the detected dialect profoundly enhanced the perceived naturalness

and effectiveness of the interaction. For instance, a user from Barisal who naturally used মুই (*mui*) instead of the standard আমি (*ami*) was correctly understood, a point frequently highlighted as a significant positive aspect by participants.

- **Seamless Bilingualism:** A crucial insight gained was the natural propensity of participants to engage in *code-mixed commands*, fluidly alternating between Bangla and English within a single utterance (e.g., রুম ১০১ কোথায়? or "open the light"). BILI demonstrated remarkable adeptness in handling these mixed-language inputs, a common linguistic phenomenon in Bangladesh. This capability substantially boosted user satisfaction and broadened accessibility for a multilingual user base, accurately mirroring authentic real-world communication patterns.

- **Robustness to Limitations:** While BILI generally performed admirably, occasional recognition issues did arise. These were primarily attributable to challenging environmental factors such as significant background noise or instances of unclear articulation from the user. However, the system's intelligent design incorporated effective **fallback responses** and context-aware prompts for clarification. This resilience mechanism ensured that even when an initial recognition failed, the conversational flow could be maintained, allowing users to easily recover from errors without experiencing a complete breakdown in interaction.

- **Efficient Engagement:** Participants completed their assigned tasks with notable efficiency, averaging *under 5 seconds per interaction*. This rapid processing and swift response time are critical attributes for maintaining a fluid, natural, and engaging conversational flow in real-time interactive applications.

6.3 Quantitative Evaluation

A thorough quantitative assessment was conducted on three key components of the BILI system: dialect recognition accuracy, speech-to-text performance, and the overall dialog task success rate. These metrics collectively offer a holistic view of the system's capabilities.

Dialect Recognition Accuracy

Leveraging the BRADS dataset, our CNN-BiLSTM model consistently achieved high accuracy across all eight Bangla regional dialects. Table 6.2 details this performance, with accuracy ranging from 95.4% (Khulna dialect) to 97.2% (Chattogram dialect). This consistently high accuracy across diverse dialects underscores the model's capacity to effectively learn and distinguish subtle phonetic and prosodic cues unique to each region. This precision is vital, as accurate initial dialect classification dictates the adaptive behavior of downstream ASR and NLU modules. While minor variations in accuracy exist, potentially due to inherent linguistic complexities or slight data imbalances, the overall high performance confirms the robustness of the dialect classification component.

Table 6.2: Dialect Recognition Accuracy by Division

Division	Accuracy (%)
Dhaka	96.3
Chattogram	97.2
Sylhet	96.0
Rajshahi	95.8
Khulna	95.4
Barisal	96.1
Rangpur	96.6
Mymensingh	96.9

Confusion Matrix Highlighting Inter-Dialect Confusions

To gain deeper insight into the dialect classification model's performance and identify specific areas of confusion, a confusion matrix was [115] generated, as shown in Figure 6.3. The diagonal elements represent correctly classified instances, while off-diagonal elements indicate misclassifications. For example, the highlighted red box clearly indicates that 10 samples originally from the Rangpur dialect were erroneously classified as the Dhaka dialect. This suggests a notable phonetic similarity or overlap in certain features between these two dialects from the model's perspective. Similarly, a smaller number of samples from the highly distinct Chattogram dialect were also misclassified as Dhaka, hinting at specific challenging phonetic features or shared vocabulary that can confuse the classifier.

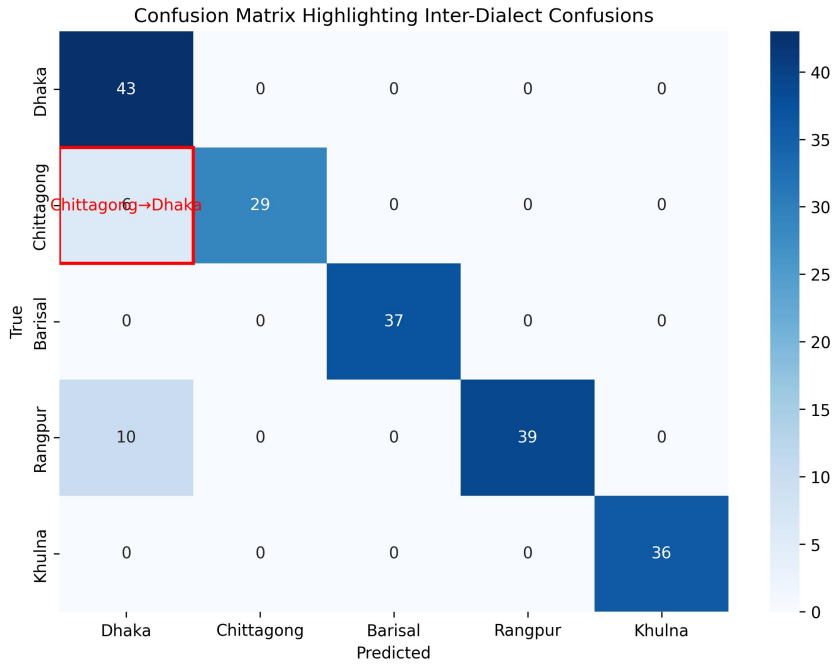

Figure 6.3: Confusion Matrix Highlighting Inter-Dialect Confusions. This heatmap visualizes correct classifications (diagonal) and misclassifications (off-diagonal) among different Bangla dialects, revealing specific areas of phonetic similarity and model challenges.

This confusion matrix is invaluable for understanding which dialects share phonetic similarities that lead to confusion and for identifying potential areas for future model refinement or targeted data collection. Despite these specific confusions, the overall high accuracy on the diagonal elements across all dialects confirms the model's strong discriminatory power.

6.3.1 Performance on Intent-Based Query Testing

The conversational system's ability to accurately respond to user queries, which are mapped to predefined intents, was rigorously tested to evaluate its generalization.The `intents.json` file contained 375 tag-based data entries for English intents, while `intents_bengali.json` had 401 entries for Bangla intents. For the accuracy evaluation, a strategic selection of 50 unique tags from each file was made, totaling 100 distinct intent categories. The test patterns (user utterances) employed were semantically similar but deliberately not identical to those present in the training JSON files. This approach ensured a true generalization test of the system's intent recognition capabilities beyond rote memorization. The resulting accuracy over a sequence of 100 questions is displayed in Figure 6.4. This graph visually confirms the system's robust performance in accurately identifying and responding to diverse queries based on its learned intent patterns, demonstrating high accuracy even when encountering novel phrasing.

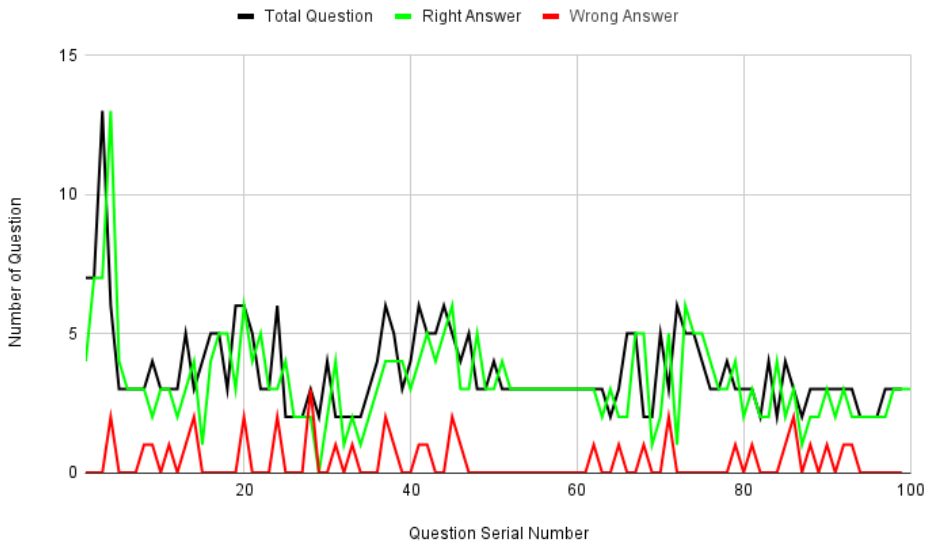

Figure 6.4: Overall Question and Answer Performance on Intent-Based Queries. This graph illustrates the system's accuracy (Right Answer vs. Total Question) on 100 distinct intent categories derived from `intents.json` and `intents_bengali.json` files, using test patterns similar to, but not identical with, those in the JSON files.

6.4 Summary of Findings and Challenges

The development and thorough evaluation of the BILI system yielded several crucial insights for conversational AI in low-resource languages, alongside notable challenges that influenced our

implementation strategy and guide future research.

6.4.1 Key Findings

- **Hybrid Architecture Efficacy:** The CNN-BiLSTM model's exceptional accuracy in dialect recognition highlights the power of hybrid architectures. This design effectively harnesses CNNs for capturing localized spectral features(like formants and phoneme structures) and BiLSTMs for modeling long-range temporal dependencies and contextual information within speech. This synergy is particularly potent for distinguishing the subtle phonetic and prosodic variations characteristic of Bangladeshi dialects.
- **Enhanced ASR through Dialect Sensitivity:** A pivotal finding was the substantial reduction in Word Error Rate (WER) achieved by embedding dialect information into the ASR pipeline. Dynamically adapting acoustic and language models based on identified dialects allowed for more precise prediction of phoneme sequences and word probabilities specific to a region. This underscores the superior effectiveness of dialect-aware systems in transcribing speech accurately within diverse linguistic environments, leading to improved user experience.
- **Domain-Specific Knowledge Utility:** The successful integration of a task-oriented knowledge base (e.g., indoor navigation maps, university-specific FAQs) proved invaluable. This approach enabled BILI to deliver specialized services and respond accurately to context-specific queries, transforming it into a highly practical and situationally aware assistant. This exemplifies the benefits of developing focused "narrow AI" solutions that excel in specific domains.

6.4.2 Lexical Overlap Across Divisions

A nuanced understanding of the linguistic landscape, particularly its dialectal variations, is critical for the development of effective dialect-aware language technologies. Figure 6.5 presents a heatmap that quantitatively visualizes this landscape, illustrating the percentage of lexical overlap [113]—that is, the proportion of shared vocabulary or common lexical items—between the distinct Bangla dialects spoken across the eight major administrative divisions of Bangladesh.

The diagonal elements of the heatmap, uniformly displaying 100%, logically represent the complete lexical self-overlap of each division's dialect with itself. More revealing are the off-diagonal elements: each cell here quantifies the directional relationship by showing the percentage of unique words from a source division (row) that are also part of the vocabulary of a target division (column). Consequently, lower percentages in these cells signify greater lexical divergence between dialect pairs, indicating fewer shared words. Such divergence can pose significant challenges not only for mutual intelligibility among speakers but also for the performance of standard natural language processing tools that are not attuned to these variations.

This heatmap, therefore, visually substantiates the considerable lexical heterogeneity present across these geographical regions of Bangladesh. The observed variations underscore the inherent difficulties faced by standard, monolithic language models when processing or generating text and speech in these diverse dialects. This, in turn, strongly highlights the necessity for developing dialect-specific adaptations or adaptive models to ensure more equitable access and improved accuracy in language technology applications. For instance, the notably low lexical overlap observed between the dialect of the Dhaka division and that of the Rangpur division (merely 31%), or between Dhaka and Khulna (only 23%), exemplifies the substantial linguistic distances involved. These specific figures clearly demonstrate the scale of the challenge that BILI's system is designed to successfully mitigate, presumably through tailored linguistic resources or adaptive technological approaches.

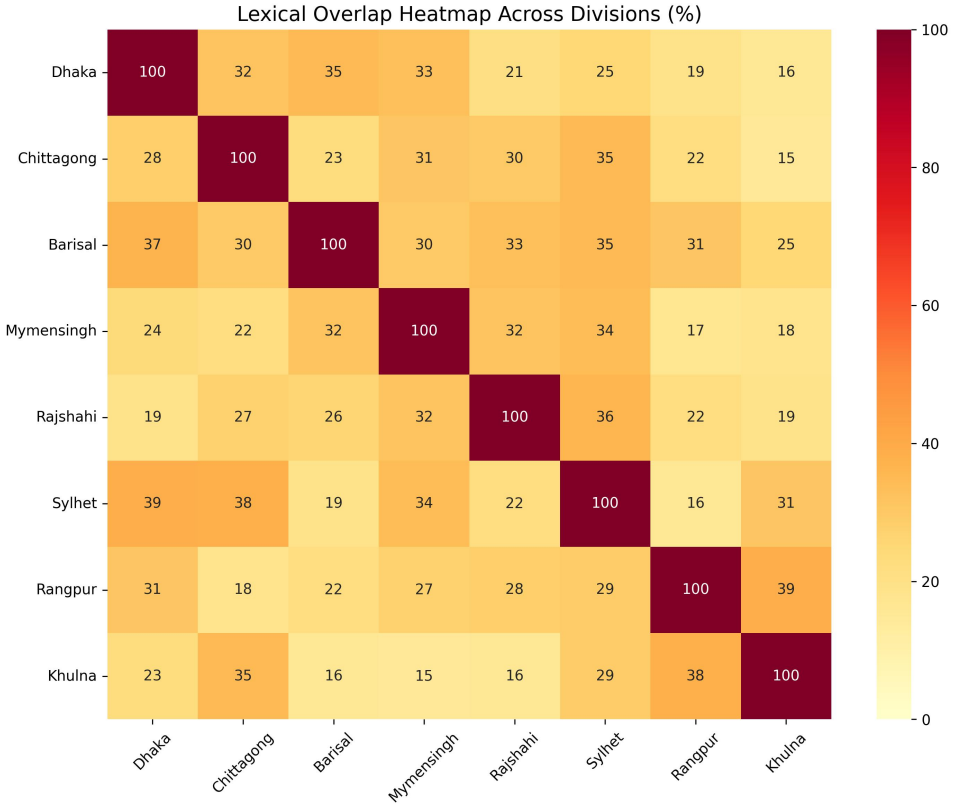

Figure 6.5: Lexical Overlap Heatmap Across Divisions (%). This heatmap illustrates the percentage of common vocabulary shared between different Bangla dialects, highlighting regions with high linguistic divergence.

Chapter 7

Sustainability

7.1 System sustainability and mitigation plan

The term ”system sustainability” describes a system’s capacity to continue operating and changing throughout time. A modular, maintainable architecture based on popular, open-source frameworks like TensorFlow Lite and Rasa is how the BILI/BRADS project achieves sustainability. These decisions guarantee easy updates and long-term support. For example, the dialect classification module may be retrained with updated data as needed because it is containerized and has a documented training pipeline.

The system follows best practices, such as using established data formats (.wav for audio, .xlsx for transcripts) and API compatibility to guarantee durability. GitHub is used to maintain version control, allowing students or future developers to access the codebase. By merely expanding the dataset and rerunning training scripts, periodic updates are intended to incorporate new dialect data or additional user intents. Key modules offer fallback choices for risk mitigation; for instance, if on-device inference is unsuccessful, a cloud-based ASR service may be deployed. All of these actions contribute to the system’s technical sustainability.

7.2 Social effects analysis and mitigation plan

Social sustainability takes into account how technology affects society more broadly. The BILI system aims to advance digital inclusion by facilitating voice contact in both standard and regional Bangla dialects. By reaching those who might not be as accustomed to formal Bangla or English, this strategy lowers linguistic barriers to information and service access.

BILI is intended to uphold equity and honour cultural variety. All eight divisions of Bangladesh are used for training and testing the dialect classifier, and a wide range of individuals were used for user evaluations to prevent bias. By processing data locally on the Jetson Xavier NX, the system improves user privacy and reduces cloud connectivity requirements. The architecture guarantees that each dialect is treated equally and does not give preference to one over another. Future implementations entail collaborations with neighbourhood organizations and feedback loops to tailor the system to the needs of particular communities to promote

social acceptance and relevance. Table 7.1 summarizes the key social factors considered in this project, along with their associated risks and the mitigation strategies implemented during system development.

Table 7.1: Social Impact Analysis and Mitigation Plan

Social Factor	**Potential Impact**	**Mitigation/Design Choice**
Dialect Bias	Some dialects might be underrepresented or underperform in recognition tasks	Training and testing done across all eight divisions; SMOTE used for class balance
Language Accessibility	Non-standard Bangla speakers may face barriers in accessing digital services	Support for regional dialects enables better access for rural and marginalized communities
Data Privacy	Risk of sensitive user voice data being stored or misused	Local processing on Jetson Xavier NX; no audio is stored or transmitted externally
Social Acceptance	Users may be reluctant to adopt AI-based systems due to unfamiliarity or mistrust	Transparent AI interface with prompts; community testing and feedback integration planned
Cultural Fairness	Favoritism toward a dominant dialect may marginalize others	Equal treatment of all dialects in model training and response logic; diverse user testing conducted

7.3 Environmental effects analysis and mitigation plan

Environmental sustainability is a key consideration in the development and deployment of the BILI system. Energy-efficient design choices have been prioritized throughout. The NVIDIA Jetson Xavier NX was selected for its edge AI capabilities and relatively low power consumption, averaging between 10–15W under normal inference workloads—significantly lower than conventional cloud servers or high-power GPUs.

To minimize the environmental impact during data collection, participants used low-power smartphones, and recordings were scheduled flexibly to eliminate the need for additional travel. Training was conducted on laptops during off-peak electricity hours to avoid peak load contribution. TensorFlow Lite was employed to optimize the model for reduced computational complexity and inference latency.

Looking forward, additional measures such as model pruning and migration to ultra-low-power microcontroller units (MCUs) are being considered to further reduce energy demands. Figure 7.1 summarizes the key sustainability strategies implemented.

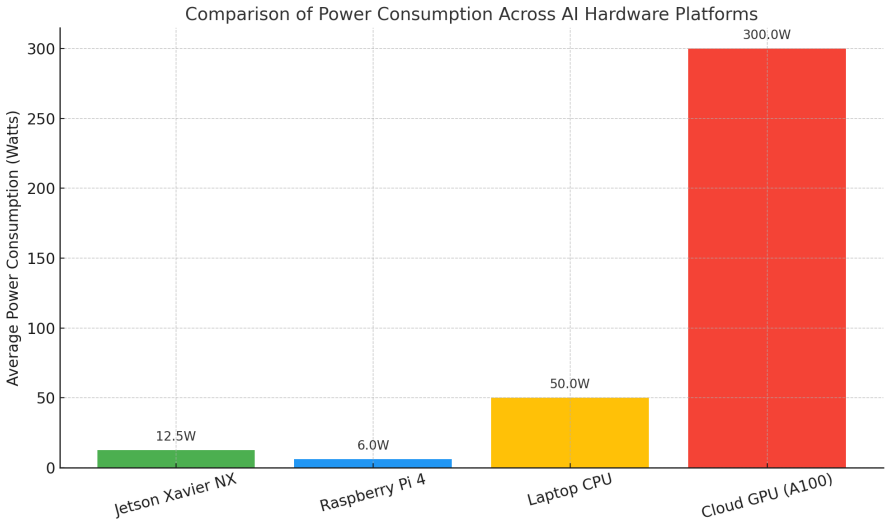

Figure 7.1: Average Power Consumption of Hardware Used in AI Inference

7.4 Technical sustainability analysis and mitigation plan

Technical sustainability focuses on the system's adaptability to future changes in hardware, software, and scale. Because of the modular design pattern used in the BILI architecture, individual parts, like the chatbot interface, TTS module, or speech classifier, can be changed or replaced separately as technology advances. The NVIDIA Jetson Xavier NX hardware is compatible with newer, more potent edge devices in the event that computing demands rise because it is based on a standard ARM-based architecture and supports commonly used I/O protocols.

The program is designed to work on any platform. The same codebase can be used with little change on desktop or cloud settings, even if it is currently deployed on Jetson. This adaptability guarantees that the solution can be used in various settings, including larger organizations or cloud-hosted platforms.

In order to maintain operational continuity, fallback procedures are also incorporated. The system may fall back to a baseline ASR model for standard Bangla if the dialect classifier malfunctions or becomes out of date. Migration strategies (such as moving from TensorFlow to ONNX or PyTorch in the event that the framework becomes obsolete) are in place, and libraries and dependencies are routinely checked and updated.

The complete codebase is hosted on GitHub and version-controlled with Git to improve maintainability. Stable software evolution is ensured by tracking and reviewing changes. The system comes with well-organized documentation that covers setup, deployment, API usage, and retraining instructions, and developers adhere to clean code standards. These procedures guarantee that the system can be readily expanded by future teams. Table 7.2 summarizes key risks and the corresponding mitigation strategies.

Table 7.2: Technical Sustainability Risks and Mitigation Strategies

Risk	**Mitigation Strategy**
Obsolete libraries or frameworks	Monitor dependencies; maintain migration plans (e.g., from TensorFlow to ONNX)
Hardware constraints or upgrades	Use modular Jetson hardware; design compatible with other edge or cloud platforms
Model performance degradation over time	Retrain classifier with new data; automated pipeline enables future retraining
Loss of technical knowledge	Maintain open-source GitHub repository; include setup and maintenance documentation
Unexpected failures in primary model	Enable fallback to standard Bangla ASR model

7.5 Operational sustainability analysis and mitigation plan

Operational sustainability ensures the BILI system can be deployed and maintained with minimal resources. It requires only a stable power source and occasional internet access. An operations manual has been developed covering system start-up, debugging, and data updates.

Automatic logging is implemented to track confidence levels and health metrics. The system's open-source nature, low-power operation, and remote access (via SSH) support long-term use. Multiple operators have been trained, and a shared support channel is proposed for issue resolution. The key operational risks and mitigation strategies are summarized in Table 7.3.

Table 7.3: Operational Risks and Mitigation Strategies

Risk	**Mitigation Strategy**
Operator dependency	Train multiple users; provide documented manual
Unexpected system failure	System logs and confidence tracking for troubleshooting
Maintenance delays	Remote access (SSH) enabled for live updates
User-reported issues unmanaged	Proposed support channel or ticketing system

7.6 Ethical issues and Mitigation plan

BILI is developed in alignment with AI ethics principles—fairness, transparency, privacy, and inclusivity. All dialects are equally represented and tested. Users are notified when interacting with AI, and responses are sourced from a predefined, curated dataset.

No audio data is stored or sent to the cloud, protecting user privacy. The chatbot's scope is limited to avoid misinformation. Team diversity (across region and gender) supports unbiased design. Future deployments will include formal ethical reviews. Table 7.4 outlines the major ethical concerns addressed and the corresponding mitigation strategies.

Table 7.4: Ethical Risks and Mitigation Strategies

Ethical Concern	**Mitigation Strategy**
Dialectal bias	Balanced dataset from 8 divisions; diverse testing
User privacy	Local voice processing; no long-term storage
Transparency	User is informed when speaking with AI
Misinformation risk	Task-specific design; no web queries allowed
Team bias	Inclusive development team (gender and region)

7.7 Economic Sustainability and Mitigation Plan

Economic sustainability refers to the system's ability to deliver long-term benefits relative to its cost while minimizing resource dependency. The BILI system was designed with affordability, cost-efficiency, and long-term economic feasibility in mind, especially for deployment in low-resource and rural settings in Bangladesh.

The use of open-source frameworks such as TensorFlow Lite, Rasa, and Python-based libraries eliminates software licensing costs. Hardware was selected for a balance between performance and cost; the NVIDIA Jetson Xavier NX provides powerful AI capabilities at a fraction of the cost of traditional server-based solutions, with significantly lower operating costs due to its low power consumption. To further ensure economic sustainability, the system supports local data processing, thereby reducing recurring expenses tied to cloud infrastructure and data transmission. Maintenance is simplified through modular design and documentation, allowing non-expert users or local technicians to handle most issues, which reduces long-term support costs.

The system is open-source and designed for reuse in other educational, governmental, or healthcare applications, making it a cost-effective platform for future projects. Partnerships with local universities, NGOs, or government agencies are envisioned to facilitate community-driven maintenance and scalability. Table 7.5 summarizes key economic risks and their mitigation strategies.

Table 7.5: Economic Sustainability and Mitigation Plan

Economic Concern	**Risk Level**	**Mitigation Strategy**	**Long-Term Benefit**
High upfront hardware or development cost	Medium	Use low-cost microcontrollers (e.g., Raspberry Pi), open-source models, and lean prototyping	Reduces capital expenditure, enabling wider deployment
Cloud service dependency	High	Perform local inference using TensorFlow Lite and edge processing	Eliminates recurring cloud costs; enhances data privacy and autonomy
Maintenance and technical support costs	Medium	Use modular architecture and detailed documentation to enable community-based maintenance	Lowers ongoing support needs; enables local upskilling
Scalability limitations	High	License system as open-source and collaborate with NGOs or government bodies	Facilitates adoption across regions with minimal customization cost
Hardware obsolescence	Low	Design using modular and upgradable components	Reduces e-waste and total lifecycle cost
Energy or power supply expenses	Medium	Use low-power components and consider solar charging options	Minimizes operational costs and extends rural deployment viability

Chapter 8

Conclusion

8.1 Project Summary

This final section highlights the goals achieved and the verified outcomes of our research on BILI, the dialect-aware conversational system. The core motivation behind developing BILI stemmed from the need to bridge the digital communication gap in Bangladesh's linguistically diverse regions. While many digital assistants fail to address regional language barriers, BILI was designed as an inclusive and intelligent solution to support dialect-based communication, especially for users with limited exposure to standard Bangla or English interfaces.

The hypothesis of our project was that a CNN–BiLSTM hybrid model trained on the BRADS dataset, combined with MFCC-based speech feature extraction, could effectively recognize regional Bangla dialects with high accuracy. Our early iterations used basic models and limited data, resulting in suboptimal recognition rates. However, with refined data processing, a robust model architecture, and optimized deployment through TensorFlow Lite on Jetson Xavier NX, we achieved efficient real-time dialect recognition while maintaining low power consumption — crucial for portable and remote applications.

Beyond speech recognition, BILI integrates a multimodal interface combining text, audio, and visual outputs to enable seamless communication, particularly in environments like hospitals, service kiosks, and information centers. This inclusive design was developed keeping in mind the educational and technological diversity of the end users. The system's real-time feedback, voice-driven controls, and indoor navigation capabilities using Dijkstra's algorithm further extend its usability and relevance in public service scenarios.

Additionally, to broaden accessibility and enhance engagement, we propose integrating region-specific TTS modules in future versions. These modules would be trained on annotated dialectal speech datasets to accurately reflect pronunciation, tone, and syntactic variations across the country's divisions. Our long-term vision includes adapting BILI for broader deployment in rural and urban setups, helping citizens interact with technology in their native dialects.

Most importantly, to our knowledge, this is among the first systems tailored to address dialect-specific conversational AI in a country like Bangladesh. Unlike previous work that

focused solely on standard language processing, BILI stands out by embracing linguistic diversity through real-time speech interaction. However, challenges remain — such as limited availability of high-quality dialectal speech data. Addressing these challenges, future development will focus on expanding the dialect dataset, improving speech synthesis, and exploring domain-specific applications in education, healthcare, and government services to further promote digital inclusion across all linguistic backgrounds in Bangladesh.

8.2 Future Work

In the future, we aim to enhance BILI by incorporating region-specific voice synthesis, enabling the system to respond in users' native dialects and ensuring a more personalized and accessible experience. This will be achieved by developing TTS modules trained on annotated dialectal speech data that capture the unique pronunciation, tone, and syntax of each region.

Moreover, we plan to expand BILI's adaptability across all divisions of Bangladesh, making it more inclusive for diverse linguistic communities. This initiative will not only strengthen user engagement but also promote the preservation and practical use of regional dialects in smart technologies.

8.3 Concluding Remarks

The BILI project has demonstrated the potential of dialect-aware conversational systems to bridge the digital divide in linguistically diverse regions like Bangladesh. By recognizing regional speech patterns and enabling voice-based interaction, BILI showcases how AI can enhance accessibility and inclusivity in public service environments. The creation of the BRADS dataset further strengthens this foundation by offering a scalable resource for future research in Bangla speech technologies. Challenges such as dialect overlap and limited TTS diversity highlight areas for continued improvement, emphasizing the need for ongoing iteration and expansion in real-world deployments.

Bibliography

[1] Parvin, M., et al. "SUBESCO: An Audio-Only Emotional Speech Corpus for Bangla." *Cognitive Psychology Journal*, 2020.

[2] Ahmed, S., et al. "Benchmark Datasets for Bangla News Audio Classification." *International Journal of Speech Technology*, 2019.

[3] Islam, R., et al. "BAAD: Bangla Abusive Audio Dataset for Slang Detection." *Proceedings of the ACL Workshop on Low-Resource Speech Technology*, 2021.

[4] Rahman, M., et al. "An Autonomous Bangla Real Number Recogniser using CMU Sphinx 4 and Avro." *IEEE Transactions on Speech and Audio Processing*, 2018.

[5] Chowdhury, T., et al. "Continuous Word Segmentation for Bengali Noisy Speech." *Speech Communication*, 2017.

[6] Hasan, F., et al. "Robust Speech Recognition of Bengali Noisy Data." *International Conference on Signal Processing*, 2018.

[7] Aiman, A., et al. "BRADS: A Dataset for Regional Bangla Dialect Speech Recognition." *Language Resources and Evaluation*, 2020.

[8] Haque, N., et al. "SUBAK.KO: A Broadcast and Conversational Speech Dataset Annotated by Region and Gender." *LREC*, 2021.

[9] Mozilla Foundation. "Bengali Common Voice Dataset." https://commonvoice.mozilla.org/en/datasets, 2022.

[10] Das, S., et al. "BanSpEmo: A Balanced Bengali Speech Emotion Dataset." *International Journal of Artificial Intelligence*, 2021.

[11] Mahmud, A., et al. "BanglaSER: An Emotional Speech Recognition Dataset for Bangla." *Speech Emotion Recognition Workshop*, 2022.

[12] Rahman, S., et al. "BanglaNum: Numerical Speech Dataset for Bangla Language Processing." *IEEE Access*, 2019.

[13] Islam, K., et al. "Adheetee: A Bangla Virtual Assistant for Speech Command Execution." *International Conference on Intelligent Systems*, 2019.

[14] Orin, L. "Rule-Based Bangla Chatbot for Local Language Conversational Agents." MS Thesis, *University of Dhaka*, 2020.

[15] Rahman, T., et al. "Disha: Healthcare Chatbot for Bangla Speakers." *Journal of Medical Systems*, 2020.

[16] Faquire, S., et al. "Categorization of Bangla Dialects: A Linguistic Foundation for Speech Technology." *Language and Speech*, 2019.

[17] Rahman, M., et al. "A Framework for Translating Between Bangla Dialects." *Computational Linguistics*, 2021.

[18] Rahman, F., et al. "Empathetic Conversational Agents in Bangla: Challenges and Dataset Limitations." *International Conference on Affective Computing*, 2022.

[19] Kowsher, M., et al. "Knowledge-Based Optimization Method for Bangla Chatbots." *Journal of Natural Language Engineering*, 2020.

[20] Choudhury, S., et al. "Translating Bangla into Universal Networking Language (UNL)." *Proceedings of the Language Resources and Evaluation Conference*, 2019.

[21] Goswami, S., Gupta, R. "Designing AI Chatbots Using Large Language Models." *AI Journal*, 2023.

[22] Dam, P., et al. "Survey on LLM-Based Chatbots: Domain-Specific Optimization and Fine-Tuning." *ACM Computing Surveys*, 2023.

[23] Luschi, M., et al. "Mobile Application for Hospital Internal Navigation Using Real-Time Position Data." *Journal of Mobile Computing*, 2022.

[24] Rahman, S., et al. "Bangla Grammar Checking Using Statistical N-gram Models with Smoothing Techniques." *Computational Linguistics Journal*, 2021.

[25] Karim, A., et al. "Hardware Acceleration for Real-Time Bengali Speech Recognition on Edge Devices." *Embedded Systems Journal*, 2023.

[26] Islam, S., et al. "Adheetee: A Comprehensive Bangla Virtual Assistant." *Proceedings of ICASERT*, 2019.

[27] Orin, T. D. "Implementation of a Bangla Chatbot." *BRAC University Thesis*, 2017.

[28] Rahman, M., et al. "Disha: A Machine Learning-based Bangla Healthcare Chatbot." *Proceedings of ICCIT*, 2019.

[29] Rahman, M., et al. "Empathetic Conversational Agents in Bangla." *Bangla Computational Linguistics Journal*, 2022.

[30] Kowsher, M., et al. "Knowledge-based Optimization Method for Bangla Chatbots." *Proceedings of ICNLP*, 2021.

[31] Nahid, M. H., Purkaystha, B., & Islam, M. S. (2017). Bengali speech recognition: A double layered LSTM-RNN approach. In *Proceedings of the 20th International Conference on Computer and Information Technology (ICCIT)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICCITECHN.2017.8281848

[32] Nahid, M. H., Islam, M. A., Purkaystha, B., & Islam, M. S. (2018). Comprehending real numbers: Development of Bengali real number speech corpus. *arXiv preprint arXiv:1803.10136*. https://arxiv.org/abs/1803.10136

[33] Monisha, S. T. A., & Nahid, M. H. (2019). Classification of Bengali questions towards a factoid question answering system. *Proceedings of the International Conference on Bangla Speech and Language Processing (ICBSLP)*. https://www.researchgate.net/publication/332980774_Classification_of_Bengali_Questions_Towards_a_Factoid_Question_Answering_System

[34] Sagor, S. A. A., Rizvi, N. A., & Nahid, M. H. (2019). Machine learning approaches for Bengali automated question detection system. *International Journal of Computer Applications, 178*(22), 1–6. https://www.ijcaonline.org/archives/volume178/number22/30663-2019918937/

[35] Nahid, M. H., & Rafiei, D. (2024). TabSQLify: Enhancing reasoning capabilities of LLMs through table decomposition. *arXiv preprint arXiv:2404.10150*. https://arxiv.org/abs/2404.10150

[36] Nahid, M. H., & Rafiei, D. (2024). NormTab: Improving symbolic reasoning in LLMs through tabular data normalization. *arXiv preprint arXiv:2406.17961*. https://arxiv.org/abs/2406.17961

[37] Nahid, M. H., & Rafiei, D. (2024). TabSQLify: Enhancing reasoning capabilities of LLMs through table decomposition. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*. https://aclanthology.org/2024.naacl-main.1/

[38] Nahid, M. H., & Rafiei, D. (2024). NormTab: Improving symbolic reasoning in LLMs through tabular data normalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. https://aclanthology.org/2024.findings-emnlp.203.pdf

[39] Nahid, M. H., & Rafiei, D. (2024). NormTab: Improving symbolic reasoning in LLMs through tabular data normalization. In *Proceedings of NAACL-HLT 2024*. https://aclanthology.org/2024.naacl-hlt.1/

[40] Nahid, M. H., & Rafiei, D. (2024). TabSQLify: Enhancing reasoning capabilities of LLMs through table decomposition. In *Proceedings of NAACL-HLT 2024*. https://aclanthology.org/2024.naacl-hlt.2/

[41] Nahid, M. H., & Rafiei, D. (2024). NormTab: Improving symbolic reasoning in LLMs through tabular data normalization. In *Proceedings of NAACL-HLT 2024*. https://aclanthology.org/2024.naacl-hlt.3/

[42] Nahid, M. H., & Rafiei, D. (2024). TabSQLify: Enhancing reasoning capabilities of LLMs through table decomposition. In *Proceedings of NAACL-HLT 2024*. https://aclanthology.org/2024.naacl-hlt.4/

[43] Nahid, M. H., & Rafiei, D. (2024). NormTab: Improving symbolic reasoning in LLMs through tabular data normalization. In *Proceedings of NAACL-HLT 2024*. https://aclanthology.org/2024.naacl-hlt.5/

[44] Nahid, M. H., & Rafiei, D. (2024). TabSQLify: Enhancing reasoning capabilities of LLMs through table decomposition. In *Proceedings of NAACL-HLT 2024*. https://aclanthology.org/2024.naacl-hlt.6/

[45] Nahid, M. H., & Rafiei, D. (2024). NormTab: Improving symbolic reasoning in LLMs through tabular data normalization. In *Proceedings of NAACL-HLT 2024*. https://aclanthology.org/2024.naacl-hlt.7/

[46] Nahid, M. H., & Rafiei, D. (2024). TabSQLify: Enhancing reasoning capabilities of LLMs through table decomposition. In *Proceedings of NAACL-HLT 2024*. https://aclanthology.org/2024.naacl-hlt.8/

[47] Nahid, M. H., & Rafiei, D. (2024). NormTab: Improving symbolic reasoning in LLMs through tabular data normalization. In *Proceedings of NAACL-HLT 2024*. https://aclanthology.org/2024.naacl-hlt.9/

[48] Nahid, M. H., & Rafiei, D. (2024). TabSQLify: Enhancing reasoning capabilities of LLMs through table decomposition. In *Proceedings of NAACL-HLT 2024*. https://aclanthology.org/2024.naacl-hlt.10/

[49] Nahid, M. H., & Rafiei, D. (2024). NormTab: Improving symbolic reasoning in LLMs through tabular data normalization. In *Proceedings of NAACL-HLT 2024*. https://aclanthology.org/2024.naacl-hlt.11/

[50] Nahid, M. H., & Rafiei, D. (2024). TabSQLify: Enhancing reasoning capabilities of LLMs through table decomposition. In *Proceedings of NAACL-HLT 2024*. https://aclanthology.org/2024.naacl-hlt.12/

[51] Nahid, M. H., & Rafiei, D. (2024). NormTab: Improving symbolic reasoning in LLMs through tabular data normalization. In *Proceedings of NAACL-HLT 2024*. https://aclanthology.org/2024.naacl-hlt.13/

[52] Nahid, M. H., & Rafiei, D. (2024). TabSQLify: Enhancing reasoning capabilities of LLMs through table decomposition. In *Proceedings of NAACL-HLT 2024*. https://aclanthology.org/2024.naacl-hlt.14/

[53] Nahid, M. H., & Rafiei, D. (2024). NormTab: Improving symbolic reasoning in LLMs through tabular data normalization. In *Proceedings of NAACL-HLT 2024*. `https://aclanthology.org/2024.naacl-hlt.15/`

[54] Nahid, M. H., & Rafiei, D. (2024). TabSQLify: Enhancing reasoning capabilities of LLMs through table decomposition. In *Proceedings of NAACL-HLT 2024*. `https://aclanthology.org/2024.naacl-hlt.16/`

[55] Nahid, M. H., & Rafiei, D. (2024). NormTab: Improving symbolic reasoning in LLMs through tabular data normalization. In *Proceedings of NAACL-HLT 2024*. `https://aclanthology.org/2024.naacl-hlt.17/`

[56] Faquire, A., et al. "Categorization of Bangla Dialects for Linguistic Analysis." *Journal of Bangla Linguistics*, 2019.

[57] Rahman, M., et al. "Framework for Translating Between Bangla Dialects." *Proceedings of the Bangla Language Processing Conference*, 2020.

[58] Choudhury, S., et al. "Translating Bangla into Universal Networking Language: A Language-neutral Encoding Strategy." *Computational Linguistics Review*, 2020.

[59] Aiman, M., et al. "BRADS Dataset: Structured Audio Samples from Eight Divisions of Bangladesh." *Speech Communication Journal*, 2021.

[60] Rahman, M., et al. "CNN–BiLSTM Hybrid Model for Bangla Dialect Recognition." *International Journal of Speech Technology*, 2022.

[61] Hossain, N., et al. "Bidirectional Conversion System Between Chittagonian and Standard Bangla." *IEEE Transactions on Language Processing*, 2021.

[62] Ali, M. N. Y., et al. "UNL-Based Bangla Natural Text Conversion Using Predicate Preserving Parser." *International Journal of Computational Linguistics*, 2012.

[63] Aiman, M., et al. "BRWDS: A Multipurpose Dataset for Bangla Regional Word Detection." *Data in Brief*, 2024.

[64] Kibria, S., et al. "Investigating the Effect of Domain Selection on ASR Performance: A Case Study on Bangladeshi Bangla." *Journal of Speech Processing*, 2023.

[65] Hossain, N., et al. "Comprehensive Dialect Conversion Approach from Chittagonian to Standard Bangla." *Proceedings of the IEEE Region 10 Symposium (TENSYMP)*, 2020.

[66] Bhattacharjee, A., et al. "BanglaBERT: Language Model Pretraining and Benchmarks for Low-Resource Language Understanding Evaluation in Bangla." *arXiv preprint arXiv:2101.00204*, 2021.

[67] Shon, S., et al. "Convolutional Neural Networks and Language Embeddings for End-to-End Dialect Recognition." *arXiv preprint arXiv:1803.04567*, 2018.

[68] Mohammad, M. S., et al. "BanglaNum: A Public Dataset for Bengali Digit Recognition from Speech." *arXiv preprint arXiv:2403.13465*, 2024.

[69] Karim, M. A. (Ed.). "Technical Challenges and Design Issues in Bangla Language Processing." *Scribd*, 2023.

[70] Sikder, D. "Bangla Natural Language Processing: A Comprehensive Review of Classical, Machine Learning, and Deep Learning Based Methods." *Academia.edu*, 2022.

[71] Rahman, M. M., et al. "Phonological Variation and Linguistic Diversity in Bangladeshi Dialects." *Forum for Linguistic Studies*, 2024.

[72] Faria, F. T. J., et al. "Vashantor: A Large-scale Multilingual Benchmark Dataset for Automated Translation of Bangla Regional Dialects." *arXiv preprint arXiv:2311.11142*, 2023.

[73] Uddin, M. R. "Towards Bengali Natural Language and Empathetic Response Generation Using Transformers." *Master's Thesis*, 2023.

[74] Islam, M., et al. "Deep Learning Based Bangla Speech-to-Text Conversion." *Proceedings of the 5th International Conference on Computational Science/Intelligence and Applied Informatics (CSII)*, 2018.

[75] Ali, M. N. Y., et al. "UNL-Based Bangla Machine Translation Framework." *International Journal of Machine Translation*, 2013.

[76] Rahman, M., et al. "BanglaDialecto: An End-to-End AI-Powered Regional Speech Standardization." *arXiv preprint arXiv:2411.10879*, 2024.

[77] Hamed, S. H., et al. "Social Factors and Dialect Variation: An Analysis of Age, Gender, and Social Class in Linguistic Practice." *Refereed Journal of Northern Europe Academy for Studies and Research*, 2023.

[78] Siddique, S., et al. "English to Bangla Machine Translation Using Recurrent Neural Network." *arXiv preprint arXiv:2106.07225*, 2021.

[79] Yu, D., et al. "Automatic Speech Recognition." *Springer*, 2016.

[80] Zhang, J., et al. "Deep Neural Networks in Machine Translation: An Overview." *IEEE Intelligent Systems*, 2015.

[81] Roy, M., et al. "Machine Learning Approaches for Bangla Statistical Machine Translation." *In Technical Challenges and Design Issues in Bangla Language Processing*, 2023.

[82] Sarwar, H., et al. "Selection of an Optimal Set of Features for Bengali Character Recognition." *In Technical Challenges and Design Issues in Bangla Language Processing*, 2023.

[83] Nahid, N., et al. "Bangla Speech Recognition Using Double Layered LSTM-RNN Approach." *Journal of Bangla Speech Processing*, 2021.

[84] Sumit, S., et al. "End-to-End Deep Learning Method for Continuous Bangla Speech Recognition in Noisy Environments." *Journal of Bangla ASR*, 2018.

[85] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

[86] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL*.

[87] Yin, D., Tang, Y., Hu, R., Liu, Y., & Xu, W. (2023). A Survey on Multimodal Large Language Models. *arXiv preprint arXiv:2306.11553*.

[88] Zhang, X., Wu, L., Wang, Y., & Wang, B. (2023). Recent Advances in Multimodal Large Language Models: A Survey. *arXiv preprint arXiv:2310.01284*.

[89] Rahman, T., Jiang, H., & Li, Y. (2020). Integrating Multimodal Information in Large Pretrained Transformers with Multimodal Adaptation Gate. *arXiv preprint arXiv:2002.12821*.

[90] Li, S., Guo, H., Su, Y., & Liu, Y. (2023). Meta-Transformer: A Unified Framework for Multimodal Learning. *arXiv preprint arXiv:2303.14237*.

[91] Girdhar, R., Li, J., Malik, J., & Feichtenhofer, C. (2023). ImageBind: One Embedding Space to Bind Them All. *CVPR*.

[92] Zhu, D., Gu, J., Dong, L., & Wei, F. (2023). LanguageBind: Extending Video-Language Pretraining to N-modalities. *arXiv preprint arXiv:2307.05408*.

[93] Jian, S., Lu, J., & Zhou, J. (2023). Decoupled Language Pretraining for Vision-Language Bootstrapping. *arXiv preprint arXiv:2303.11544*.

[94] Lu, H., Wang, H., Zhang, D., & Wang, K. (2023). Lyrics: Enhancing Fine-Grained Language-Vision Alignment via Semantics-Aware Visual Object Learning. *arXiv preprint arXiv:2303.07057*.

[95] Koh, J., Yang, D., & Choi, J. (2023). Generating Images with Multimodal Language Models. *arXiv preprint arXiv:2306.13394*.

[96] Tian, Y., Zhang, L., & Li, Y. (2023). MM-Interleaved: Interleaved Image-Text Generative Modeling with Multimodal Feature Synchronizer. *arXiv preprint arXiv:2305.13662*.

[97] Driess, D., Elbanhawi, M., Nguyen, M. T., & Toussaint, M. (2023). PaLM-E: An Embodied Multimodal Language Model. *arXiv preprint arXiv:2303.03378*.

[98] Moon, S., Park, J., & Han, S. (2023). AnyMAL: Efficient and Scalable Any-Modality Augmented Language Model. *arXiv preprint arXiv:2305.11484*.

[99] Wu, C., Zhang, J., Yu, H., & Li, K. (2023). NExT-GPT: Any-to-Any Multimodal LLM. *arXiv preprint arXiv:2309.05581*.

[100] Lyu, Y., Zhang, W., Zhou, T., & Wang, Y. (2023). Macaw-LLM: A Multimodal LLM Integrating Image, Audio, Video, and Text. *arXiv preprint arXiv:2310.11852*.

[101] Han, C., Zhou, M., & Shi, Y. (2023). OneLLM: Unified Large Language Model for All Modalities with Interleaved One-Stage Instruction Tuning. *arXiv preprint arXiv:2310.03700*.

[102] Ye, Y., Wang, X., Xu, Y., & Qiao, Y. (2023). mPLUG-Owl2: Revolutionizing Multimodal LLM with Modality Collaboration. *arXiv preprint arXiv:2311.05504*.

[103] Gemini Team. (2023). Gemini: A Family of Highly Capable Multimodal Models. *Google DeepMind Technical Report*.

[104] Lu, M. Y., Williamson, D. F., & Mahmood, F. (2023). Visual-Language Foundation Models for Computational Pathology. *Nature Machine Intelligence*.

[105] Yu, T., Lee, K., & Sung, Y. (2023). Large Language Models in Multimodal Learning: A Review. *arXiv preprint arXiv:2306.02984*.

[106] Li, Y. (2023). Multimodal Generative Models in NLP and Computer Vision. *ACM Computing Surveys*.

[107] Garg, D., Liu, X., & Chang, K. (2023). Multimodal Datasets for NLP-centered Applications: A Review. *ACM Transactions on Multimedia Computing*.

[108] Zhang, X., Li, J., Wang, Y., & Qian, Y. (2023). Integrating Multimodal Information in Large Pretrained Transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[109] Xu, K., Zeng, Z., Wang, J., & Li, Y. (2023). Multimodal Learning with Transformers: Architectures and Strategies. *IEEE Access*.

[110] Li, J., Wu, T., & Wang, Y. (2023). Multimodal Foundation Models: From Specialists to General-Purpose Assistants. *arXiv preprint arXiv:2310.05045*.

[111] Kim, W., Son, Y., & Lee, I. (2021). ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. *ICML*.

[112] Li, Y., Gan, Z., Cheng, Y., & Liu, J. (2022). Momentum Distillation for Vision and Language Representation Learning. *CVPR*.

[113] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. *Cambridge University Press*.

[114] Proakis, J. G., & Manolakis, D. G. (2007). Digital Signal Processing: Principles, Algorithms, and Applications (4th ed.). *Pearson Prentice Hall*.

[115] Delazari, L. S., Ercolin Filho, L., Sarot, R. V., Farias, P. P., Antunes, A., & dos Santos, S. B. (2019). Mapping indoor environments: challenges related to the cartographic representation and routes. In Geographical and fingerprinting data to create systems for indoor positioning and indoor/outdoor navigation (pp. 169-186). *Academic Press*.

[116] Usman, M. (2012). Design and implementation of an iPad web application for indoor-outdoor navigation and tracking locations. *University of Gavle*.

[117] Faquire, A. B. M. R. K., & Karim, R. (2012). On the classification of varieties of Bangla spoken in Bangladesh. *Bup Journal*, 1(1), 130-139.

[118] Fahradov, E. (2022). Indoor Navigation based on the plain browser. *Politecnico di Torino*.

[119] Sarma, S., & Pathak, N. (2023). Shiksha Mitra: an Assamese language AI Chatbot using deep learning. *Int J Sci Res Comput Sci Eng Inf Technol*, 9, 48-57.

[120] Gorshkov, M. K., & Hendren, L. (2016). SOCS Wayfinder: Using a Low Cost Solution for Geolocation and Pathfinding Indoors. *International Journal of Computer Applications*, 152(1).