CLASSIFICATION OF BANGLADESHI REGIONAL LANGUAGE USING MACHINE LEARNING AND DEEP LEARNING

BY

SAKHAOYAT ULLAH TANVIR
ID: 203-15-14471

AND

YEASIN MIA
ID: 203-15-14543

FINAL YEAR DESIGN PROJECT REPORT

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

Supervised By

Mr. Saiful Islam
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised By

Mr. Amir Sohel
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

June 2024

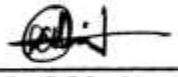
APPROVAL

This Project titled "Classification Of Bangladeshi Regional Language Using Machine Learning And Deep Learning", submitted by **SAKHAOYAT ULLAH TANVIR** and **YEASIN MIA** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 14-07-2024.

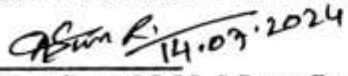
BOARD OF EXAMINERS

Dr. S.M Aminul Haque — **Board Chairman**
Professor & Associate Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Md. Abbas Ali Khan — **Internal Examiner 1**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University



Mr. Md. Aynul Hasan Nahid — **Internal Examiner 2**
Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University



Dr. Abu Sayed Md. Mostafizur Rahaman — **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Mr. Saiful Islam, Assistant Professor, Department of Computer Science and Engineering, Daffodil International University**. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

13/07/24

Mr. Saiful Islam
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

13/7/24

Mr. Amir Sohel
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

Submitted by:

Tanvir

Sakhaoyat Ullah Tanvir
ID: 203-15-14471
Department of CSE
Daffodil International University

Yasin

Yeasin Mia
ID: 203-15-14543
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty for His divine blessing making it possible for us to complete the final year project/internship successfully.

We are grateful and wish our profound indebtedness to **Mr. Saiful Islam**, **Assistant Professor** Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Machine Learning*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

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

We would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

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

ABSTRACT

The goal of this project is to reliably detect linguistic variants and dialects by classifying regional languages spoken in Bangladesh using machine learning (ML) and deep learning (DL) techniques. The dataset has 3000 entries, with a sufficient representation of each of the five major regional languages (Chattogram: 655, Dhaka: 608, Rangpur: 621, Sylhet: 553, Noakhali: 562). The entries are distributed among these five major languages. The procedure of collecting data included developing a survey form, obtaining and preparing text samples, and cleaning data using natural language processing methods. Neural Bayes (BNB), Support Vector Machines (SVM), Random Forest, Bi-directional Long Short-Term Memory (Bi-LSTM), Logistic Regression (LR), and Convolutional Neural Networks (CNN) were among the ML and DL models that were assessed. According to the results, DL models (Bi-LSTM: 95.24%, CNN: 98.48%) are much better at classifying regional languages than classic ML methods (Random Forest: 70.00%, SVM: 67.78%, LR: 66.22%, BNB: 64.44%). All in all, this study highlights how well DL methods capture complex linguistic patterns that are essential for problems involving the classification of regional languages. It highlights the importance of Bangladesh's language diversity from a cultural standpoint and promotes ethical research methods to help preserve languages and promote social inclusion. Prospective avenues for investigation encompass augmenting the intricacy of the model through syntactic and semantic evaluations, in addition to examining the wider sociocultural implications of language categorization technology.

.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	ii
Declaration	iii
Acknowledgements	iv
Abstract	v
Table of contents	vi-viii
List of Figures	viii-ix
List of Tables	x
CHAPTER	
CHAPTER 1: INTRODUCTION	**1-13**
1.1 Introduction	1
1.2 Motivation	7
1.3 Rationale of the Study	7
1.4 Research Questions	8
1.5 Expected Output	10
1.6 Project Management and Finance	11
1.7 Report layout	13
CHAPTER 2: BACKGROUND	**26-24**
2.1 Preliminaries/Terminologies	16
2.2 Related Works	16

2.3 Comparative Analysis and Summary 20

2.4 Scope of the Problem 21

2.5 Challenges 22

2.6 Summary 24

CHAPTER 3: RESEARCH METHODOLOGY 25-32

3.1 Research Subject and Instrumentation 25

3.2 Data Collection Procedure/Dataset Utilized 26

3.3 Statistical Analysis 28

3.4 Proposed Methodology/Applied Mechanism 28

3.5 Implementation Requirements 32

CHAPTER 4: EXPERIMENTAL RESULT 34-45

4.1 Experimental Setup 34

4.2 Experimental Results & Analysis 34

4.3 Discussion 45

CHAPTER 5: IMPACT ON SOCIETTY 47-50

5.1 Impact on Society 47

5.2 Impact on Environment 58

5.3 Ethical Aspects 49

5.4 Sustainability Plan 50

CHAPTER 6: SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH **50-51**

6.1 Summary of the Study **51**

6.2 Conclusions **51**

6.3 Implication for Further Study **52**

REFERENCES **54-55**

APPENDIX **56-61**

LIST OF FIGURES

FIGURES	PAGE NO
Figure 1.1: Representation of Bernoulli Naive Bayes (BNB) [18]	2
Figure 1.2: Proposed SVM Architecture(SVM) [19]	3
Figure 1.3: Logistic Regression (LR) Architecture [20]	4
Figure 1.4: Architecture of Random Forest [21]	4
Figure 1.5: Architecture of Bidirectional Long Short-Term Memory (Bi-LSTM) [22]	5
Figure 1.6: Representation of the Architecture of a Bidirectional Long Short-Term Memory (CNN) [23]	6
Figure 3.1: Overview of Datasets	27
Figure 3.2: Working Flow.	29
Figure 3.3: Some Raw data	30
Figure 3.4: Data Labeling	31
Figure 4.1: Model Accuracy and Loss curve of (CNN Networks).	39
Figure 4.2: Model Accuracy and Loss curve of (Bi-LSTM)	40
Figure 4.3: Confusion Matrix (BNB)	41
Figure 4.4: Confusion Matrix (SVM)	42
Figure 4.5: Confusion Matrix (LR)	43
Figure 4.6: Confusion Matrix (RF)	44
Figure 4.7 Confusion Matrix (CNN)	44
Figure 4.8: Final Output	45

LIST OF TABLES

TABLES	PAGE NO
Table 1.1: Estimated Cost	12
Table 2.1: Comparative analysis with previous work	20
Table 3.1: Number of all datasets.	27
Table 4.1: Accuracy for Classification of Individual all model	35
Table 4.2: Precision, Recall, F1-Score, and Support (n) for BNB	36
Table 4.3: Precision, Recall, F1-Score, and Support (n) for SVM	36
Table 4.4: Precision, Recall, F1-Score, and Support (n) for LR	37
Table 4.5: Precision, Recall, F1-Score, and Support (n) for Random Forest	37
Table 4.6: Precision, Recall, F1-Score, and Support (n) Bi-LSTM	38
Table 4.7: Precision, Recall, F1-Score, and Support (n) for CNN	38

CHAPTER 1
Introduction

1.1 Introduction

Bangladesh has a wide variety of local languages that each represent the distinct identity of the nation, thanks to its rich cultural legacy and linguistic diversity. The linguistic landscape of Bangladesh is greatly influenced by these languages, which range from the lively rhythms of Bengali to the unique varieties of e Dhakaiya, Noakhailla, Rongpuri, Sylheti, and Chatgaiya. Though these regional languages are important, there is still a need for efficient ways to categorize and evaluate them, especially given the fast development of technology and the increasing significance of natural language processing (NLP) technologies.

This research is to investigate the classification of regional languages spoken in Bangladesh using both deep learning and conventional machine learning techniques in response to this demand. Through the application of contemporary computational methods, our goal is to create models that can precisely recognize and classify text samples according to the language regions in which they are spoken. This project has enormous potential for a number of uses, such as linguistic study, artistic and cultural sustainability, and language preservation.

There are many potential and obstacles associated in classifying Bangladeshi regional languages. On the one hand, it's a difficult and multifaceted undertaking because Bangladesh's linguistic diversity includes a large variety of languages, accents, and linguistic variances. However, new developments in deep learning and machine learning techniques present encouraging paths forward for overcoming these obstacles and deriving valuable insights from language data.By using computational tools, we hope to contribute to a wider awareness and understanding of Bangladeshi indigenous languages through this initiative. In light of Bangladesh's varied linguistic landscape, we seek to open the door for improved language identification, dialect detection, and regional text analysis by creating strong classification models.

We will examine the processes used in this research, including collecting data, preprocessing, selecting models, training, assessment, and testing, in the parts that follow. In the realm of linguistic studies and natural language processing, our systematic method aims to produce accurate and dependable categorization findings that can help practitioners and scholars similarly.

Bernoulli Naive Bayes (BNB):

The probabilistic classifier Bernoulli Naive Bayes (BNB) is well-known for its effectiveness and ease of use. Designed for binary-valued features, this Naive Bayes version assumes independence between the features. BNB determines the likelihood of every class given a collection of input features by utilizing the Bayes theorem. It excels at handling huge feature spaces with grace, especially with high-dimensional data. Its dependence on the feature independence assumption, however, has multiple disadvantages because real-world data frequently breaks from it. Furthermore, BNB might be sensitive to unimportant qualities, which could affect how well it predicts the future. Despite these disadvantages, its scalability and easy of use make it a popular option for a variety of machine learning tasks.

Figure 1.1: Representation of Bernoulli Naive Bayes (BNB) [18]

Support Vector Machine (SVM):

Support Vector Machine (SVM) is a powerful technique for supervised learning that may

be used for both regression and classification applications. What makes it so effective is that it finds the hyperplane on the feature space that best divides classes. SVM performs well in high-dimensional settings and provides flexibility by using different kernel functions that can handle decision boundaries that are not linear. Specifically, by determining decision boundaries using just support vectors, it preserves memory efficiency. However, careful kernel selection and parameter tweaking are necessary for the efficient use of SVM, and its computing needs may present difficulties when working with huge datasets.

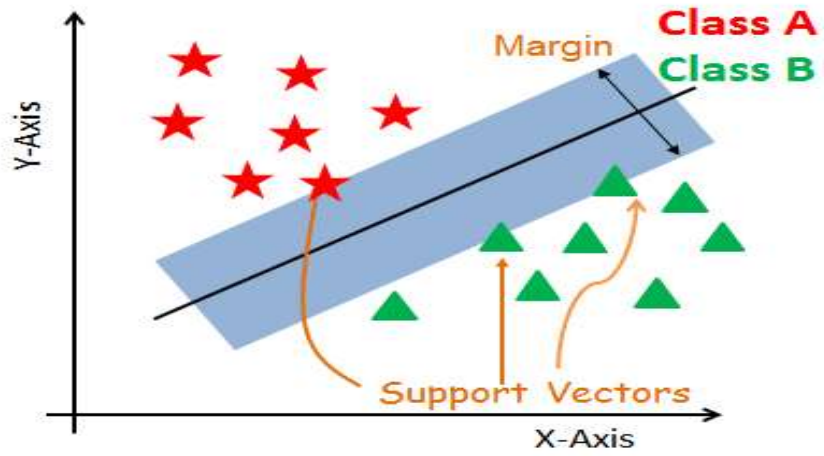


Figure 1.2: Proposed SVM Architecture(SVM) [19]

Logistic Regression (LR):

Based on the logistic function, logistic regression (LR) is a fundamental framework for binary and multiple classes classification tasks that calculates the likelihood that an input is part of a particular class. LR provides effective training and prediction procedures and is praised for being straightforward and understandable. Decision-making skills are improved by its capacity to produce probability results. LR's effectiveness may be constrained, nevertheless, by its dependence on the premise of a linear connection among characteristics and the log- odds of the result, especially in situations when the data contains nonlinear interactions. This limitation notwithstanding, LR is still a useful and popular technique, particularly in situations where understanding and computational performance are important concerns.

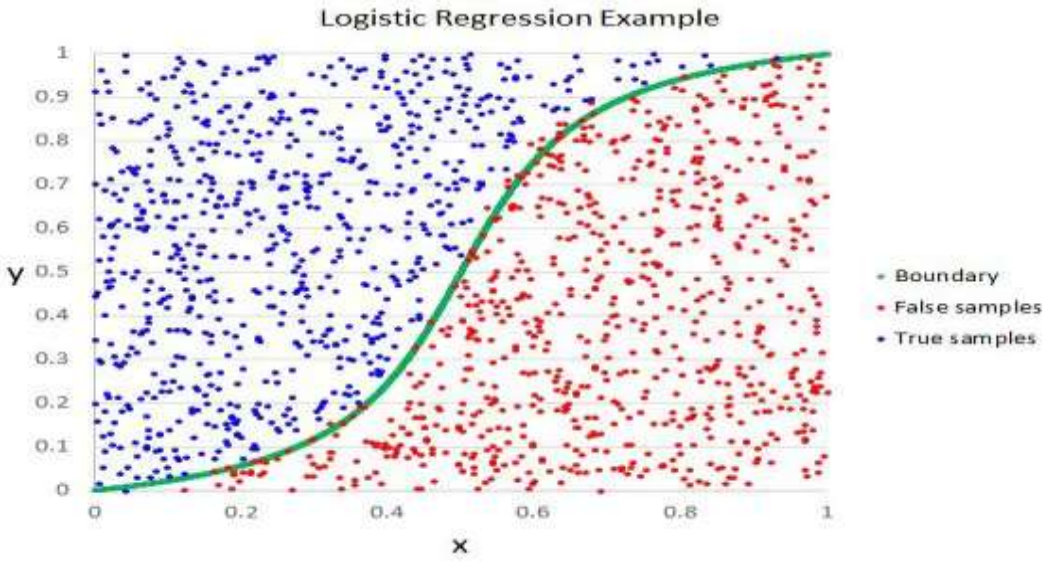


Figure 1.3: Logistic Regression (LR) Architecture [20]

Random Forest:

Multiple decision trees are built during training using Random Forest, a group learning technique, and their predictions are combined for classification tasks.

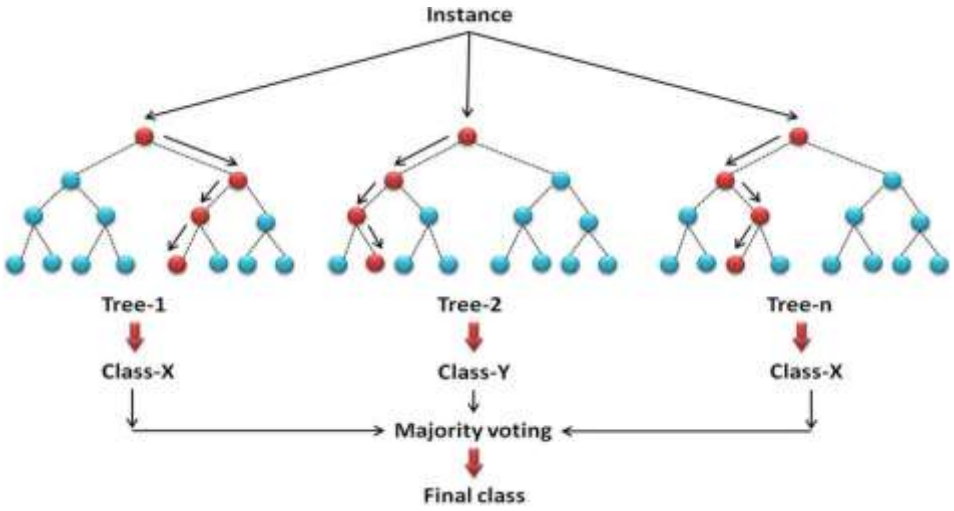


Figure 1.4: Architecture of Random Forest [21]

Because it uses several weak learners, it is strong against overfitting, which is its main strength. Random Forest provides insights into feature significance and is effective at managing high-dimensional data. With big datasets or a lot of trees, it might, however,

show slower prediction speeds. Furthermore, as comparison to individual decision trees, its interpretability is slightly weakened. Random Forest is still a popular tool for classification jobs because of its stability and capacity to manage complicated data well, even in light of these factors.

Bidirectional Long Short-Term Memory (Bi-LSTM):

Bidirectional Long Short-Term Memory (Bi-LSTM) is an effective recurrent neural network design that can process input sequences in both directions, allowing it to store context from both the past and the future. Bi-LSTM is mostly applied to sequence labeling applications such as language modeling as well as sentiment analysis. It is good at learning complex patterns, capturing long-term dependencies in sequential data, as well as effectively managing variable-length input sequences. But the main disadvantages are that it is computationally expensive, especially when dealing with lengthy sequences, and it can have problems with vanishing or expanding gradients when training. Because Bi-LSTM can effectively represent complex sequential data, it continues to be a popular option despite these difficulties.

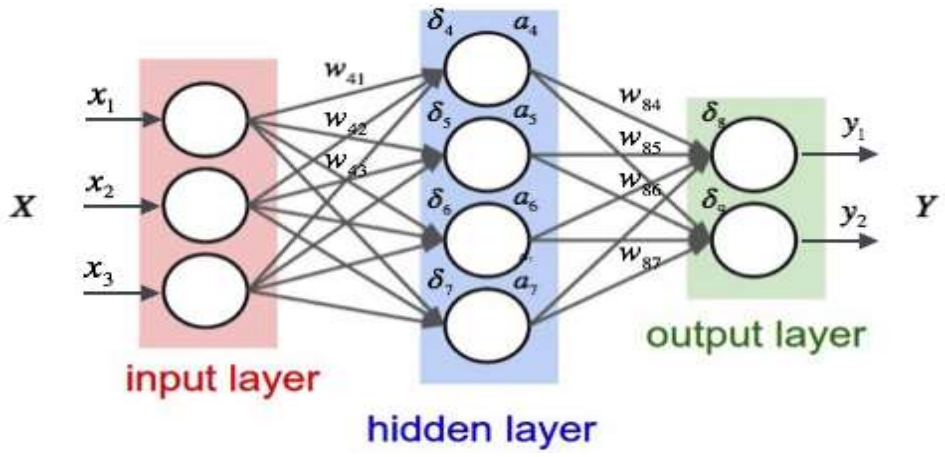


Figure 1.5: Architecture of Bidirectional Long Short-Term Memory (Bi-LSTM) [22]

Convolutional Neural Network (CNN):

Convolutional Neural Net (CNN), although well recognized for their abilities in analyzing images, are flexible models that may be used to sequential data, such as text, by employing methods like 1D convolutions. With convolutional, collecting, and

completely connected layers, CNNs are particularly good at identifying local patterns in data. This is made possible by parameter sharing, which lowers the amount of parameters that need to be learned. A certain level of translation consistency is also provided by them. However, in comparison with Recurrent Neural Network (RNN), CNN may require a significant quantity of data for efficient training and may be less skilled at identifying long-range relationships in sequential data. Despite these weaknesses, CNNs are still essential in a variety of fields because of their prowess in feature extraction and pattern detection.

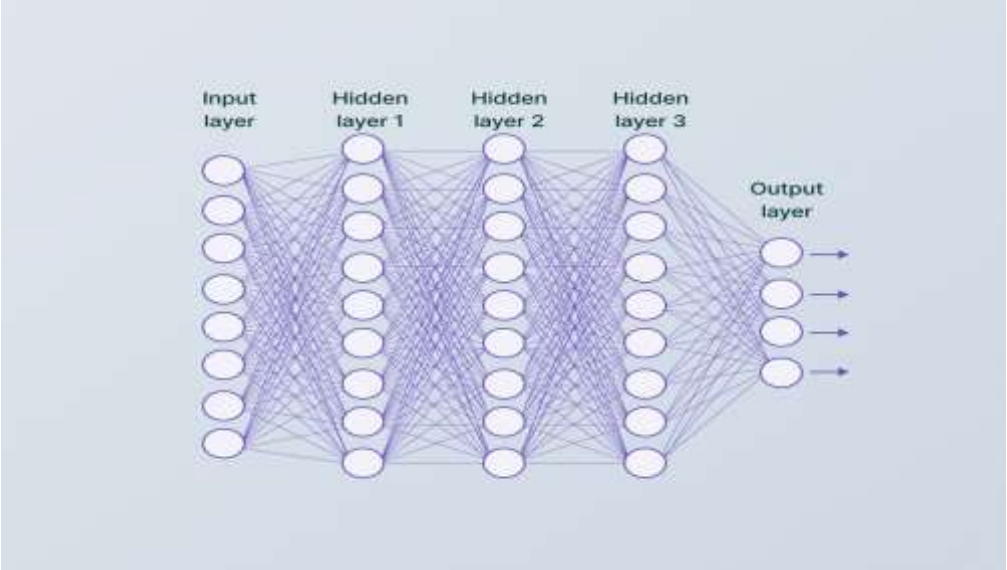


Figure 1.6: Representation of the Architecture of a Bidirectional Long Short-Term Memory (CNN) [23]

Natural Language Processing (NLP)

Transition learning is a method that improves performance in a different context by utilizing knowledge from an initial data set in a single task or field. This method adds new layers that are taught for the particular target job while discarding the last few elements of a pre-trained network. Task-specific attributes are captured by the first layers of the deep learning process. In essence, transfer learning is using knowledge from one environment to enhance competence in a different Raghu et al. (2019). We investigate CNN topologies with the goal of finding a model that can be used for applied learning in

picture categorization. Experimenting and adjusting network topology settings and dataset features can assist find aspects influencing classification success, even with

limited processing power and time.

Underperforming deep learning models might result from training them on a little dataset. This problem is addressed via transfer learning, wherein Neyshabur et al. (2020) initialize a model with weights from another model learned on a bigger, unrelated dataset. Notable gains are especially noticeable when handling a more intricate segmentation task and a smaller target training sample.

To summarize, the process of transfer learning entails initializing a model using weights derived from training an unrelated, larger dataset. This leads to notable performance gains, particularly in situations when segmentation tasks are difficult and there is a limited amount of target training data.

1.2 Motivation

The project's motivation comes from the inherent importance of protecting and comprehending Bangladesh's linguistic diversity. Regional languages of Bangladesh are not only means of communication but also reflect distinct social identities, historical legacies, and cultural subtleties. However, because there are so many diverse dialects, variants, and linguistic characteristics found in different places, it is extremely difficult to classify and analyze these languages.

We hope to overcome these issues and improve our proficiency in Bangladeshi local languages by utilizing methods from deep learning and machine learning through this project. Many benefits stem from being able to precisely categorize and evaluate various languages, such as improved communication, cultural legacy preservation, diversity of languages, and support for educational endeavors.

Additionally, we want to provide researchers, linguists, teachers, and legislators with useful instruments for dialect identification, regional text analysis, and language preservation through the development of strong categorization models. Our ultimate goal is to use technology to its fullest potential in the fields of linguistic studies and natural language processing, while simultaneously advancing the wider understanding and conservation of Bangladeshi linguistic heritage.

1.3 Rational of the Study

This research was motivated by the realization that Bangladesh, a nation known for its

rich history and diverse language environment, has a vast amount of cultural and linguistic diversity. Although Bengali is the official language, many other regional languages and dialects are many other local languages and dialects spoken in Bangladesh, each having distinct traits and historical significance. Nevertheless, there is a vacuum in the literature and study concerning the categorization and examination of these regional languages, in spite of their significance.

The goal of this project is to close this gap by creating models that are accurate for the categorization of regional languages spoken in Bangladesh using the capabilities of deep learning and machine learning techniques. We hope to accomplish this in order to support the preservation, comprehension, and enjoyment of Bangladesh's linguistic legacy. The study additionally attempts to investigate the possible uses of these models in a number of fields, such as linguistic research, cultural heritage conservation, language preservation, and educational programs.

Moreover, the research is in line with the overarching objectives of enhancing the domain of natural language processing, or NLP, and encouraging multidisciplinary cooperation among language researchers, computer scientists, and scholars. Our goal is to gain new insights into the foundation, variation, and development of the regional languages spoken in Bangladesh by fusing computational methodologies with linguistic expertise. This will improve our comprehension of the diversity of languages and advance the conversation on language and culture.

1.4 Research Questions

- **How well can Bangladeshi regional languages be classified using textual data using machine learning and deep learning techniques?**

The categorization findings of Bangladeshi regional languages using machine learning and deep learning approaches are promising. Deep learning models like Bi-LSTM and CNN are excellent at catching intricate language nuances, which results in more precise classification outcomes than more conventional techniques like SVM, LR, and RF. The precise classification of regional languages spoken in Bangladesh could be greatlyenhanced by combining these methods with extensive datasets and efficient feature engineering.

- **Which linguistic traits and qualities are most important in categorizing Bangladeshi regional languages?**

The classification of Bangladeshi regional languages relies heavily on phonologically, morphological, and syntactic characteristics. A few significant characteristics that set these languages apart include their vowel-consonant supplies, tone, grammatical procedures, word order, and subtle semantic differences. For accurate language classification, one must comprehend certain linguistic characteristics.

- **How do various deep learning and machine learning algorithms stack up in terms of performance and accuracy for language categorization tasks?**

Deep learning algorithms, such as CNN and Bi-LSTM, are better at capturing intricate linguistic patterns than classic machine learning techniques like SVM and LR, which is why they typically do better on language categorization tasks. Nonetheless, data sets and task difficulty have an impact on each algorithm's performance.

- **What effect do linguistic subtleties and dialectal differences have on the precision of language categorization models?**

Precision in language categorization models is impacted by linguistic intricacies and dialectal variations. Phonological, morphological, syntactic, and semantic variations make classification difficult and may result in incorrect classifications. Taking these subtleties into consideration is crucial to increasing model accuracy.

- **How do the quantity and diversity of the dataset impact the classification models' performance and capacity for generalization?**

Performance and generalization of classification models are significantly impacted by the quantity and diversity of datasets. Models can learn strong patterns and perform successful generalization with larger and more varied datasets. Smaller or less varied datasets, on the other hand, can result in overfitting and restricted generalization. Increasing the amount and diversity of the dataset is therefore essential to enhancing the model's performance and cross-linguistic applicability.

- **Can regional languages and dialects in Bangladesh that are closely related be distinguished with accuracy using the developed categorization models?**

Bangladeshi regional language varieties that are closely related can be distinguished with varying degrees of accuracy using the existing categorization models. Even while they catch certain subtleties, their similarities provide problems. Accuracy might be improved by adjusting and adding features unique to each language.

- **What factors affect the discrepancies in classification findings between Bangladesh's various regions?**

The disparities in classification results between Bangladesh's regions are caused by variances in linguistic characteristics, cultural customs, historical influences, and the quality of the data in each location. Comprehending these elements is essential for precise language categorization.

1.5 Expected Output

This research project will extensively examine and classify Bangladeshi local languages using a range of deep learning and machine learning models. The project attempts to effectively categorize text samples into their respective regional languages by the application of models including the use of Support Vector Machines, Logistic Regression, Naive Bayes, Random Forest, Bidirectionally Long Short-Term Memory networks, and Convolutional Neural Networks.

Classification Accuracy:

We expect to use machine learning and deep learning algorithms to achieve high accuracy in classification rates for regional languages spoken in Bangladesh after the study is finished. It is anticipated that the models would successfully categorize text examples into the appropriate linguistic regions, proving the efficacy of the presented approaches.

Identifying Linguistic Elements:

The study aims to identify the fundamental linguistic characteristics and attributes that are significant in the classification of the various dialects spoken in Bangladesh. We aim to gain more insight into the minute linguistic variances and patterns that differentiate

different language areas by analyzing element significance levels and model judgments.

Evaluation of Model Performance:

The performance of each model (SVM, LR, BNB, Random Forest, Bi-LSTM, and CNN) will be evaluated using the following common classification metrics: accuracy, precision, recall, F1-score, the area beneath the ROC curve (AUC-ROC), and others.

Algorithm Comparative Analysis:

We hope to assess the effectiveness of various algorithms using deep learning and machine learning for language categorization tasks through a comparison analysis. We evaluate many measures like precision, recall, and F1-score in an effort to identify the best algorithms for correctly categorizing regional languages spoken in Bangladesh.

Diversity and Size of the Dataset:

The study will look into how the diversity and amount of the dataset impact the classification models' performance and capacity for generalization. Larger and more varied datasets should show increases in accuracy and robustness, highlighting the significance of representativeness and quality in dataset construction.

Distinguishing Between Languages That Are Nearly Related:

Our goal is to evaluate the classification models' capacity to distinguish between Bangladesh's distinct languages and dialects that are closely related. We hope to find obstacles and restrictions in differentiating across language regions that are closely related by looking at findings from classification and confusion matrices.

Regional Differences in the Classification Outcomes:

Lastly, the study will examine the variables affecting the differences in classification findings between Bangladesh's various areas. We seek to clarify the variables influencing language variety in the nation by looking at regional variations in linguistic traits and usage patterns.

1.6 Project Management and Finance

The successful completion of the research on the categorization of regional languages spoken in Bangladesh depends on efficient project management. The project will take an

organized approach, with tasks, deadlines, and milestones all clearly stated. To guarantee the timely delivery of deliverables, an initiative manager will supervise the tracking of progress, allocation of resources, and coordination of operations. Regular teamwork and communication will also be given top priority in order to promote efficient productivity and handle any issues that may come up throughout the project.

The project's financial requirements include setting aside money for a number of costs, such as personnel charges, publication fees, computer hardware and software resources, and data collection. To predict the amount of money required for each project phase, an extensive financial plan will be created. The project's goals will also be supported by attempts to investigate possible financing sources, such as funding for research or sponsorship opportunities. Throughout the study, transparent financial management procedures will be used to guarantee accountability and maximize the use of available resources.

Table 1.1: Estimated Cost

SN	Components	Estimated Cost (BDT)
01.	Hardware	500-1000
02.	Software and Tools	700-1200
03.	Data Collection and Processing	9,000-12000
04.	Documentation and Report Writing	3000-4000
05.	Miscellaneous	1000-2000
06.	Contingency	1500-2200
Total Estimated Cost		15,700-22,400

Table 1.1 shows a calculation of the costs associated with various project components,

expressed in Bangladeshi Taka. The table shows six major components: hardware, software and tools, data collection and processing, documentation and report writing, miscellaneous expenses, and a contingency fund. The estimated cost of hardware ranges from 500 to 1000 BDT, while software and tools fall within the same range. Data collection and processing costs are expected to be higher, ranging between 9000 and 12000 BDT, reflecting the importance and resource-intensive nature of this aspect. Documentation and report writing costs are low, ranging from 3000 to 4000 BDT. General expenses, including various additional costs, are budgeted between 500 and 1000 BDT. A contingency fund of 1500 to 2200 BDT is set aside to cover unexpected expenses or changes that arise during the project's execution. The total cost of all components is estimated to be between 15700 and 22,400 BDT, providing a comprehensive overview of the project's financial requirements.

1.7 Report layout

The suggested report has a thorough format that walks readers through the methodology, conclusions, and ramifications of the study. Every chapter has a distinct function and adds to the document's overall coherence and depth.

Chapter 1: Introduction

Explains the goals and significance of the research and gives a summary of the study on the classification of regional languages spoken in Bangladesh using machine learning and deep learning techniques.

Chapter 2: Background

Analyzes the linguistic variety and cultural importance of the regional languages spoken in Bangladesh, emphasizing the necessity for efficient classification and analysis techniques in light of the development of linguistic study and technology.

Chapter 3: Data Collection

Discusses how to gather text samples from several regional languages spoken in Bangladesh. It also covers how to create a survey form, send it to participants, and organize the results into a structured dataset.

Chapter 4: Data Preprocessing

Explains the procedures for preparing the gathered dataset, such as normalization, tokenization, text cleaning, and other data-cleaning tasks unique to natural language processing.

Chapter 5: Research Methodology

Describes the technique used in the research, including the steps involved in choosing a model, training it, assessing it, and comparing machine learning with deep learning algorithms.

Chapter 6: Experimental Result and Discussion

Demonstrates the accuracy rates, performance indicators, and analysis of the classification results that were produced from the training classification models in the experiments. examines and contrasts the performance of several algorithms while talking about the ramifications of the findings.

Chapter 7: Impact on Society, Environment

Analyzes how the study might affect Bangladeshi society, culture, and language preservation initiatives. examines the wider effects of precise language classification on linguistic variety, cultural heritage preservation, and educational programs.

Chapter 8: Summary, Conclusion, Recommendation, and Implications for Future Research

This study investigates the use of deep learning and machine learning methods in the classification of regional languages spoken in Bangladesh. When models such as SVM, LR, BNB, RF, Bi-LSTM, and CNN are compared, deep learning architectures perform better. Issues such as limited data availability and language heterogeneity are noted. More varied datasets, more complex deep learning architectures, and domain adaptation strategies should be the key goals of future research. Accurate categorization models for languages with limited resources can be developed more effectively by work with linguistic and language experts.

Given their capacity to capture intricate linguistic patterns, deep learning models—particularly Bi-LSTM and CNN—perform exceptionally well in the classification of regional languages spoken in Bangladesh. Ensemble approaches have the potential to increase accuracy as well. There are still issues, though, which suggests more study is required.

Chapter 9: References

Provides a list of all the references used in the report, including books, journals, academic papers, and other pertinent works that were consulted to support the methodology and research conclusions. provides a thorough list of references so that readers can learn more about the subject and ensures transparency and trustworthiness.

Explores how accurate heart disease prediction may affect the health care system, outcomes of patients, and resource provision.

CHAPTER 2
Background

2.1 Preliminaries/Terminologies

It is important to lay a strong foundation through preparatory exercises before beginning the study of identifying Bangladeshi regional languages. A comprehensive literature review is one of these, with a special focus on studies pertaining to Bangladesh's regional languages and dialects, in order to comprehend the body of research already done in the subject of language classification. Further insights into the variety of linguistic features and subtleties found in Bangladeshi local languages will come from early conversations with linguists and stakeholders. Additionally, to properly support the research activities, the required infrastructure—such as software environments, computing resources, and data collection tools—must be set up. The project may move on with ease and efficiency toward its goals of creating precise language model classifications for regional languages spoken in Bangladesh by laying the foundation through these preparatory stages.

2.2 Related Works

Some Ahmed, Md Faisal, et al.[4] addresses the growing need for efficient online harassment detection in Bengali, which is the seventh most spoken language in the world and is increasingly used by its speakers on online platforms. The binary classification model showed excellent results for detecting situations of bullying with an accuracy of 87.91%. The authors complied with the neural network with an ensemble technique for multiclass classification, yielding an accuracy of 85%. This study aids in the development of Bengali- language tools that may successfully combat online harassment.

Shafin, Minhajul Abedin, et al. [5] addressed information relating to the standard of products and services available via the internet as well as the possibility for fraud in the market for online retailers. The study uses classification algorithms like KNN, Decision Tree, Support Vector Machine (SVM), Random Forest, and Logistic Regression as well as sentiment analysis. Notably, SVM outperformed all other algorithms, achieving the

highest accuracy of 88.81%, showing its efficacy in evaluating customer sentiment in Bengali online shopping reviews.

Sen, Ovishake, et al. [6] analyzed pertinent research papers, the paper's main objective is to explore past, present, and future trends in Bangla Natural Language Processing (BNLP). The review includes articles that were published between 1999 and 2021, with a sizable portion (50%) appearing after 2015. The paper also addresses the current BNLP limitations and upcoming trends while discussing various approaches, including Classical, Machine Learning, and Deep Learning, along with various datasets.

Ahmed, Md Tofael, et al.[7] focused on the early detection and avoidance of online bullying in texts written in both Bangla and Romanized Bangla, with a large portion of the texts coming from YouTube video comments. Notably, Multinomial Naive Bayes outperformed others with accuracy scores of 84% for the second dataset and 80% for the combined third dataset, demonstrating efficient cyberbullying detection capabilities. Support Vector Machine (SVM) achieved the highest accuracy score of 76% for the first dataset.

Rahman, Md Mahbubur, et al. [8] addressed categorizing Bangla text documents utilizing attention-based or modern transformer models. ELECTRA outperformed BERT in terms of Recall for all cases, achieving the highest Recall value of 94.36% for the Prothom Alo dataset. Additionally, ELECTRA showed improved accuracy (96.39%) and F1 scores (94.18%) in the majority of scenarios, highlighting its efficiency in classifying Bangla text.

Rahman, Md Mahbubur, et al. [9] addressed the need for comprehensive study of Bangla documents, given the diversity of the language and sentiments related to it. The study uses data from three sources to test and validate the models, with LSTM on the Prothom Alo dataset producing the best classification accuracy of 95.42%. The paper also offers a comparison of the two models, emphasizing the success of the character-level strategy in achieving greater classification accuracy.

Chowdhury, Rumman Rashid, et al. [10] introduced a sentiment analysis procedure made for examining movie reviews written in Bangla, providing automation for figuring out how viewers feel about movies or TV shows. Additionally, the model had an accuracy of

82.42% thanks to the (LSTM) network. The performance of the model is also examined through a comparison with other machine learning techniques provided in the paper.

Rahman, Saifur, et al. [11] Used a Deep Recurrent Neural Network, this paper introduces a novel method for categorizing Bangla news documents. It is remarkable that the Deep Recurrent Neural Network with Bidirectional Long Short Term Memory (BiLSTM) achieves a remarkable accuracy of 98.33%. This accuracy for classifying Bangla text is higher than that of other well-known classification algorithms, demonstrating the potency of the suggested strategy.

Tabassum, Nusrath, et al. [12] focused on Natural Language Processing (NLP), which paves the way for computers to process enormous amounts of data. In particular, it focuses on sentiment analysis, an NLP application that evaluates sentiments or opinions, whether they are positive or negative. The research uses a Random Forest Classifier to measure the overall positivity and negativity of a document or sentence, taking into account features like unigrams, POS tagging, negation handling, and the classifier itself.

Siddique, Sumaya, et al. [13] focuses on using deep learning methods and a Jetson Nano edge device to create an automatic detection system for Bangla Sign Language (BSL). The detection accuracy of Detectron2 was the highest, with a mAP@.5 of 95% and an AP of 55%. mAP@.5 accuracy values from 85% to 97% and mAP@.5-.95 values from 41% to 53% were shown by YOLOv7. Due to its effectiveness in terms of training time and frame rate, YOLOv7 Tiny was chosen to detect Bangla sign language in real-time on the Jetson Nano edge device. The system aims to provide people in Bangladesh who have speech and hearing impairments with an affordable and practical solution.

Uddin, Abdul Hasib, et al. [14] This paper addresses synchronization errors between two sources and explores pilot-aided channel estimation for two-way relay networks (TWRNs). The effectiveness of the proposed estimation algorithms is shown to outperform that of the current channel estimation techniques, effectively reducing the negative effects of asynchronous transmissions in TWRNs.

Barua, Adrita, et al. [15] addressed the nascent stage of automatic Bengali news text classification and focused on sports news classification in Bengali. The experimental results on the test dataset reveal that the Support Vector Classifier (SVC) with a feature

space encompassing unigrams, bigrams, and trigrams achieved the highest weighted f1-score of 97.60% among all classifiers and feature combinations, demonstrating its effectiveness in Bengali sports news classification.

Neri, Federico, et al. [16] compared the sentiment toward Rai, an Italian public broadcasting service, and La7, a more vibrant private company, using more than 1000 Facebook posts about newscasts as the study's sample. Results from their tests consistently show that Recall and Precision are above 87% and 93%, respectively.

Ahmed, Md Tofael, et al. [17] focuses on the creation of a cyberbullying detection model that uses a combination of Machine Learning and Deep Learning algorithms to analyze texts in both Bangla and Romanized Bangla. Notably, the Deep Learning algorithm (CNN) outperformed others for the Bangla dataset, achieving an accuracy of 84%, while for the other two datasets, the Multinomial Naive Bayes algorithm outperformed others, achieving an accuracy of 80% for the combined dataset and 84% for Romanized Bangla.

Ain, Qurat Tul, et al. [18] emphasize the significance of sentiment analysis in social media, categorizing sentiments as positive or negative. They address the challenge of limited labeled data in NLP and review how deep learning methods, like deep neural networks and convolutional neural networks, enhance sentiment analysis. These models tackle issues such as sentiment classification, cross-linguistic concerns, and product review analysis by leveraging their automatic learning capabilities.

Alam, Tanvirul, et al. [19]investigated the use of deep learning algorithms—a departure from typical methods—for identifying Bangla text using Transformer neural network architecture. For tasks including sentiment analysis, emotion recognition, news classification, and authorship attribution, the study improved multilingual Transformer models. The models beat earlier approaches with accuracy increases of 5% to 29%, achieving state-of-the-art results on six test data sets and proving the usefulness of Transformer algorithms for Bangla text classification.

Bhowmik, Nitish Ranjan, et al. [20]offered the Bangla Text Sentiment Score (BTSC) method for polarity extraction and an extended lexical data dictionary (LDD) for Deep Learning (DL) models for Bangla text sentiment analysis. Vectorization of words using models such as Word2Vec is part of preprocessing. The research looks at different deep

learning models with fine-tuning methods, and it identifies three models that perform well for sentiment assessment tasks in Bangla text: HAN-LSTM, D-CAPSNET-Bi-LSTM, and BERT-LSTM.

2.3 Comparative Analysis and Summary

Several different methods and applications are being studied currently in Bengali language analysis and sentiment categorization. Some studies focus on specific difficulties, such as cyberbullying and online harassment detection, while others examine broader problems, such as sentiment analysis related to product evaluations and news classification. The methods range from state-of-the-art deep learning architectures like Optimus and recurrent neural networks to traditional machine learning algorithms like SVM and Naive Bayes, each with different levels of accuracy and effectiveness when dealing with Bangla text data. This research addresses a range of difficulties and paves the way for more sophisticated language comprehension and analysis in this linguistic context. Despite differences in topic and methodology, they collectively contribute to the progress of processing natural languages in the Bangla language.

Table 2.1. Comparative analysis with previous work

SL No	Author Name (with reference)	Used Algorithm	Best Accuracy with Algorithm
1	Ahmed, Md Faisal, et al. [4]	Hybrid Neural Network	87.91% (binary), 85% (multiclass)
2	Shafin, Minhajul Abedin, et al. [5]	Support Vector Machine (SVM)	88.81%
3	Ahmed, Md Tofael, et al. [7]	Multinomial Naive Bayes, Support Vector Machine (SVM)	84% (Bangla), 76% (Romanized Bangla)
4	Rahman, Md Mahbubur, et al. [8]	BERT, ELECTRA	96.39% (ELECTRA)

5	Rahman, Md Mahbubur, et al. [9]	CNN, LSTM	95.42% (LSTM)
6	Chowdhury, Rumman Rashid, et al. [10]	SVM, LSTM	88.90% (SVM), 82.42% (LSTM)
7	Ahmed, Md Tofael, et al. [17]	Multinomial Naive Bay,(CNN)	84.00%(CNN)
8	Siddique, Sumaya, et al.[13]	(BSL)	95%
9	Barua, Adrita, et al. [15]	(SVC)	97.60%
10	**Proposed Approach**	**CNN**	**98.48%**

2.4 Scope of the Problem

Applying deep learning and machine learning techniques to classify local languages spoken in Bangladesh while preserving linguistic diversity and cultural legacy is part of the problem's scope. To do this, it is necessary to discover and analyze the linguistic characteristics unique to each language region as well as to design and evaluate models for classification that can accurately distinguish between different regional languages and dialects. The challenge also involves examining the effect of dialectal variations, database size, and local differences on classification results in order to obtain insights that can direct language preservation efforts, artistic and social sustainability, and educational programs.

This study's primary focus is on the classification of regional languages spoken in

Bangladesh, with a particular emphasis on Chatgaiya, Sylheti, Rajshahi, and Dhakaiya. The project's scope includes employing survey forms to collect data, preparing written content using techniques from the processing of natural language, and training and evaluating machine learning and deep learning models for the classification of languages. The study's goal is to evaluate the efficacy of various strategies, including traditional machine learning algorithms like Random Forest, Logistic Regression, and Support Vector Machine, as well as deep learning architectures like Convolutional Neural Long Short-Term Memory. Rather than using linguistic analysis or dialectical shifts within individual languages, the study is dedicated to appropriately categorizing various regional languages that are widely spoken in Bangladesh. The study of computational approaches to language classification is also included in the scope. These include support vector models (SVM), logistic regression (LR), and random forest, as well as deep learning frameworks like convolutional neural networks (CNN) and bidirectionally long-short-term memory (Bi-LSTM). By employing these techniques, the study hopes to address the difficulties in classifying languages within the Bangladeshi language context and make greater contributions to the field's understanding of linguistic diversity and cultural identity.

2.5 Challenges

- **Linguistic Diversity:**

The great linguistic diversity that exists in Bangladesh makes it difficult to categorize Bangladeshi regional languages. Accurately capturing and reflecting the range of local languages, accents, and linguistic variances presents a substantial problem for categorization algorithms.

- **Dialectal Variations:**

The classification procedure becomes even more challenging due to the existence of dialectal variances. It can be difficult for models to distinguish between dialects that are within a single language region because of their slight variations in vocabulary, syntax, and sound.

- **Limited Resources:**

The scarcity of linguistic resources, such as linguistic corpora, annotated datasets, and language resources unique to Bangladeshi regional languages, is another difficulty. The lack of these resources can make it more difficult to train and evaluate models, which lowers the general accuracy and dependability of classification models.

- **Data Sparsity:**

Data sparsity problems make it difficult to get enough and representative information for every language region. Because there may be less text samples available in some linguistic regions, it may be more challenging to train reliable models and resulting in imbalanced datasets, which may have an impact on model performance.

- **Cultural Sensitivity:**

Tasks involving language categorization need to handle the subtleties and cultural sensitivities that come with linguistic diversity. Regional language classification necessitates a thorough comprehension of social dynamics, cultural settings, and heritage, which complicates the creation and interpretation of categorization models.

- **Generalization:**

Another problem is ensuring that classification models are generalizable across various linguistic contexts and use circumstances. Models developed using certain datasets can have biases towards particular linguistic regions or find it difficult to generalize to new data, which would limit their usefulness in real-world scenarios.

- **Computational Complexity:**

It can be computationally demanding to implement and train sophisticated machine learning and deep learning techniques for language categorization problems, necessitating a large investment of time and knowledge. Achieving a balance between computational efficiency and model complexity is crucial for creating scalable and useful categorization solutions.

2.6 Summary

The study uses deep learning and machine learning techniques to classify indigenous languages spoken in Bangladesh, including Chatgaiya, Sylheti, Rajshahi, and Dhakaiya. It highlights how crucial it is to carry out preparatory activities, such as literature reviews, stakeholder meetings, and setting up the required infrastructure. Using algorithms like SVM, LSTM, and BERT, other related papers have investigated emotion analysis, cyberbullying notice, and text classification with differing degrees of accuracy. Linguistic variety, dialectal variances, resource scarcity, data sparsity, sensitivity to culture, generalization, and computational challenge are among the difficulties. In order to address these issues, the study uses efficient categorization models to improve knowledge of linguistic variety and cultural identity.

CHAPTER 3

Research Methodology

3.1 Overview

The classification of regional languages spoken in Bangladesh using deep learning and machine learning techniques is the study's focus. The study specifically attempts to create and assess classification models based on textual data samples that can reliably differentiate between several linguistic regions, such as Bengali, Chittagonian, Sylheti, Rohingya, and others. The research's instrumentation consists of machine training and deep learning models that are implemented and trained using programming languages, software libraries, and computational tools. For model development, this comprises well-known libraries like TensorFlow, Keras, and sci-kit-learn, in addition to languages for programming like Python for tasks including processing data, model training, and evaluation.

In addition, language resources and datasets tailored to regional languages spoken in Bangladesh are essential to carrying out the study. These resources could include lexicons, linguistic corpora, annotations to data sets of text samples from various language regions, and language-specific preprocessing and analysis tools.

Additionally, text samples from individuals representing different locations and linguistic backgrounds in Bangladesh may be gathered for the study through the use of survey forms or other data gathering tools. These technologies make it possible to gather representative and varied datasets for the purpose of testing and training classification models.

The research apparatus comprises computational methods, linguistic assets, and data collection equipment to enable the creation and assessment of categorization models for regional languages spoken in Bangladesh. The project intends to aid in the knowledge and preservation of Bangladesh's linguistic diversity through its incorporation of these elements

Instrumentation:

This study's instrumentation is a complex system that combines data-gathering devices, language resources, and computational tools. To meet the computational needs of machine learning and deep learning activities, computational resources include having access to cloud platforms or high-performance computing (HPC) clusters. It is imperative to have proficiency in programming languages like Python and to be familiar with the necessary libraries and frameworks for model construction and evaluation, such as TensorFlow, Keras, scikit-learn, and PyTorch. Preprocessing and analyzing data is made easier by linguistic resources including linguistic corpora, annotated datasets, and language-specific tools.One tool used to collect data is a survey form with text sample fields to collect responses from participants in different regional languages spoken in Bangladesh.The models in this work for pneumonia diagnosis are the original BNB, SVM, LR, Random forest, Bi-LSTM, and CNN architectures. In order to provide transparent documentation of methods and results, additional documentation methods like Jupyter Notebooks, Markdown, and LaTeX are used. Maintaining participant confidentiality and the safety of data during the research process is ensured by adherence to ethical rules. In order to perform thorough research on the categorization of Bangladeshi local languages and make a significant contribution to the fields of language computation and cultural preservation, the study wants to integrate these many tools.

3.2 Data Collection Procedure

Data Creating an extensive survey form to collect text excerpts from participants who represent various regions of Bangladesh is the first step in the data collecting process for this project. Text input, regional recognition, and demographic data are all fields on the form. Distributed via social media, colleges and universities, online platforms, and community organizations, it reaches a wide range of participants. Participants submit content that reflects the use of language in daily life when instructions are clear. Tokenization, stopword elimination, and normalization are among the preprocessing steps performed on the obtained samples. A dataset of 3000 entries is produced by labeling the data using regional information after it has been anonymized to preserve

privacy: Chattogram (656), Dhaka (621), Rangpur (608), Sylhet (553), and Noakhali (562). A representative and superior dataset is guaranteed by this methodical methodology, which is used to train and assess language classification algorithms.

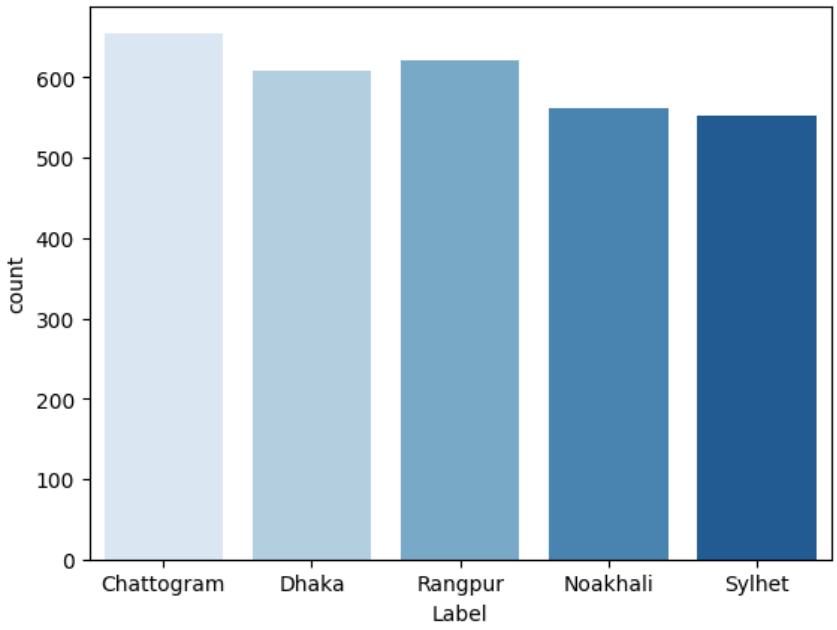


Figure 3.1: overview of Datasets

Figure 3.1: The dataset contains 3000 entries and 5 features for bangla regional language classification. A bar chart visually represents the distribution of text samples across five regional languages in Bangladesh: Chattogram, Dhaka, Rangpur, Noakhali, and Sylhet. The chart's varied blue tones emphasize the dataset's balance, ensuring sufficient representation of each regional language in training and assessment.

TABLE 3.1: Number of all datasets.

	Chattogram	Dhaka	Rangpur	Sylhet	Noakhali
Raw Dataset	655	608	621	553	562

3.3 Statistical Analysis

In order to determine whether the classification models are statistically and performance-analyzed well enough to categorize Bangladeshi regional languages, a number of indicators are evaluated. Confusion matrix analysis, accuracy, precision, recall, and F1-score are important performance indicators.

As the percentage of accurately categorized occurrences out of the total, accuracy gauges how accurate the classification models are overall. The precision of a model is determined by calculating the percentage of correctly identified occurrences among those projected to belong to a particular class. This indicates the model's ability to prevent false positives. Recall, which is another name for sensitivity, quantifies the percentage of correctly classified examples among those that actually fall into a given class, demonstrating the model's capacity to identify all pertinent cases. A more thorough evaluation of model performance is provided by the F1-score, which balances accuracy with recall within a single metric.

Confusion matrix research also offers insights on the particular kinds of errors, including misclassifications and false positives/negatives, that the classification algorithms make. The models' performance can be specifically improved by looking for patterns of misclassification in the confusion matrix. Moreover, one may use statistical methods like cross-validation and hypothesis testing to confirm the relevance of performance variations between various models or algorithms. This guarantees the categorization results' dependability and generalizability in a variety of linguistic settings and use circumstances.

Overall, the performance analysis and statistical data offer insightful information on how well the classification models differentiate between the many regional languages spoken in Bangladesh. In order to determine the best algorithms and techniques for linguistic classification challenges within the Bangladeshi language environment, the study will evaluate multiple performance measures and carry out a thorough statistical analysis.

3.4 Proposed Methodology

The proposed framework of this study aims at the determination of the Bangla regional

language through the use of the machine learning and Deep learning algorithm. The methodology includes the following steps:

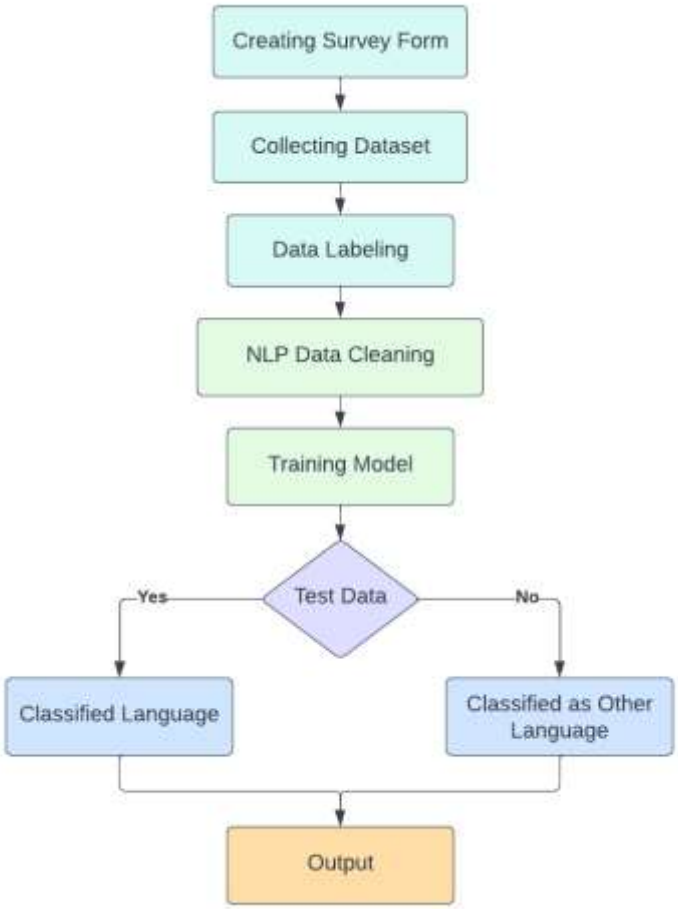


Figure 3.2: Working Flow

Data Collection:

- **Social Media Data Collection:**

Determine which pertinent social media sites—like Facebook, Twitter, and Instagram—are popular in Bangladesh.

To find postings and conversations in different dialects and linguistic regions, use hashtags and keyword searches associated with the regional languages spoken in

Bangladesh. Extrapolate text data that is accessible to the public and written in many

regional languages, such as messages, postings, and comments. Filter the gathered information to make sure it is real and relevant by eliminating spam, pointless content, and duplicate entries.To preserve user privacy and guarantee adherence to data protection laws, anonymize and combine the social media data that has been gathered.

- **Google Survey Form**:

Create a Google survey form with the express purpose of gathering text samples from respondents with varying language and geographic origins in Bangladesh.

Include metadata such as area, dialect, and demographic data in places where participants can enter text samples in their native tongues. Distribute the survey form across a range of channels, including social media, email lists, online discussion boards, and neighborhood associations, to ensure a representative and diverse sample of respondents.

To increase response rates and engagement, use incentives to encourage involvement, like gift cards or vouchers. Gather answers over a predetermined time frame, keeping an eye out for new submissions to the survey form and attending to any technical questions or concerns that respondents may have.

বলাই এ খার মুক ফাইছে ?	Sylhet
হাকাউ আইতোত আসি এটে বাল ফেলায়ছেন!	Rangpur
তোর জামাইর তইলে মাতা খারাফ অই যাইব।	Noakhali
তো দুইজ্জে লাইন্দে হনে? আজিয়ে বেক্কুনেরে উফইস্...	Chattogram
কই থাকি অউ বাতাস আয়	Sylhet

Figure 3.3: Some Raw data

There Data Validation and Quality Assurance:

To guarantee accuracy, completeness, and dependability, thoroughly validate and perform quality assurance tests on the gathered data.By comparing information from

other sources and evaluating the content's credibility, you may confirm the legitimacy of social media data Verify survey answers to make sure text samples are representative of the targeted language regions and dialects, pertinent, and coherent.Using human review and data purification techniques, resolve any disparities or inconsistencies found during the validation process.

Data Aggregation and Storage:

Create a single dataset by combining the verified information from the survey and social media platforms, then structuring it into forms that can be examined.

To ensure data integrity, security, and accessibility for further analysis, store the gathered data in compliant and secure data repositories. Throughout the research process, follow the right data management protocols to preserve the accuracy and provenance of the dataset. These include version control, backup plans, and documentation.

The study has to collect an extensive and comprehensive dataset of text samples from regional languages spoken in Bangladesh through the use of this data-collecting process. This will allow the creation and assessment of models of classification for linguistic assessment and language preservation initiatives.

Data Labeling: Using a manual labeling system, classify each text sample in the dataset by matching regional language. In order to facilitate supervised learning, this stage guarantees the generation of a labeled dataset.

Description	Label	tokenized_review
পরের লাইন আই লেহি দির	0	[পরের, লাইন, আই, লেহি, দির]
হন হারণে ভালোবাসার দাম ন দিলা	0	[হন, হারণে, ভালোবাসার, দাম, ন, দিলা]
আর বন্ধু সাদিয়া কালাইয়্যা	0	[আর, বন্ধু, সাদিয়া, কালাইয়্যা]
তোয়ার পোস্ট লাগেরদে ভালা	0	[তোয়ার, পোস্ট, লাগেরদে, ভালা]
ফুস্কা হাইয়্যুম দে চলো দুইজন	0	[ফুস্কা, হাইয়্যুম, দে, চলো, দুইজন]

Figure 3.4: Data Labeling

NLP Data Cleaning: Applying natural language processing (NLP) techniques, preprocess the labeled dataset. Tokenization, lowercasing, stop word, spelling and grammar, and special character removal are among the tasks involved. Normalizing the text data also involves stemming or lemmatization.

Training Model: Divide the preprocessed dataset into sets for validation and training. For language classification, use suitable deep learning and machine learning models, such as SVM, CNN, Random Forest, or LSTM. Utilizing the validation set to adjust hyperparameters and the training data, train the chosen models to maximize performance.

Test Data: The trained models' ability to correctly identify regional languages using a different test dataset. Determine criteria like recall, accuracy, precision, and F1-score to assess each model's efficacy. These procedures will help us create and assess classification models that will properly identify and differentiate between the various regional languages spoken in Bangladesh.

The overall goal of the data-gathering process is to compile a representative varied dataset of text samples from regional languages spoken in Bangladesh to facilitate further analysis and the creation of models for linguistic classification tasks.

3.5 Implementation Requirements

Using Computing Resources: The suggested methodology cannot be implemented without sufficient computing resources. This involves having access to cloud computing platforms or high-performance computer (HPC) clusters to meet the computational needs of deep learning and machine learning model evaluation and training.

Programming Language and Libraries: To execute the suggested methodology, one must be proficient in programming languages like Python. Model building and experimentation also require familiarity with pertinent machines as well as deep learning libraries and frameworks, including TensorFlow, Keras, scikit-learn, as well as PyTorch.

Data Collection Tools: In order to compile the dataset, survey form design tools and participant text sample collection tools are required. This could entail creating data

gathering tools specifically designed to meet the needs of the study or using survey platforms like Google Forms.

Text Processing and NLP technologies: Preprocessing and cleaning the gathered dataset require accessibility to texts processing and natural languages processing (NLP) technologies. For text processing activities, this may contain libraries like NLTK (Natural Language Toolkit), and for more complex NLP tasks like tagging parts of speech and named entity recognition, it could include spaCy.

Deep Learning and Machine Learning Frameworks: TensorFlow, Keras, scikit-learn, as well as PyTorch are just a few of the frameworks that need to be mastered in order to implement machine learning as well as deep learning algorithms. These frameworks offer programming interfaces (APIs) for creating, honing, and assessing classification models with a range of architectures and techniques.

Data Visualization Tools: Model performance metrics, classification results, and dataset statistics can all be seen with the use of data visualization tools like Matplotlib, Seaborn, and Plotly. Analyzing patterns, trends, as well as conclusions from the data and models is made easier with the help of effective data visualization.

Documentation and Version Control: Jupyter Notebooks, Markdown, as well as LaTeX are just a few of the documentation tools that are necessary to document the

implementation process, which includes code, methods, and conclusions. Git and other version control systems improve collaborative development and change tracking during the implementation phase.

Ethical Considerations: Throughout the implementation process, respect to legal requirements and data privacy laws is crucial. It is imperative to implement protocols that guarantee participant confidentiality, data integrity, and ethical management of private data gathered throughout the research.

The study can successfully implement the suggested technique and achieve its goals of creating and assessing classification models for regional languages spoken in Bangladesh by meeting these implementation requirements.

CHAPTER 4
Experimental

4.1 Experimental Setup

The suggested methodology is being tested in an experimental setting that makes use of a stable computer environment furnished with top-notch hardware and software. Access to computational resources that can meet the needs of training and assessing deep learning and machine learning models for linguistic classification tasks is part of the setup. Among these are GPU-accelerated computing resources, including GPUs manufactured by NVIDIA (Graphics Processing Units), which, in comparison to conventional CPU-based systems, significantly speed up model training and inference operations. Furthermore, for effective processing of data and model training, access to multiple component Cpu (Central Processing Units), sufficient quantities of random access memory (RAM), and fast storage are necessary. A full range of program tools and frameworks for ML, deep learning, and processing natural languages applications are also included in the experimental configuration. Using a range of techniques and architectures, well-known frameworks like TensorFlow, Keras, scikit-learn, as well as PyTorch offer APIs for generating, honing, and assessing categorization models. To further aid in the presentation of dataset research, model performance indicators, and classification outcomes, the system contains data visualization tools like Matplotlib, Seaborn, and Plotly. Transparency, reproducibility, and ethical research practices are ensured throughout the implementation phase of the experimental setup by including version control systems, efficient documentation tools, and ethical concerns. The study aims to contribute to advances in linguistic research and natural language processing by providing an efficient experimental setup that would enable thorough testing, accurate analysis, and trustworthy assessment of model classification for regional languages spoken in Bangladesh.

4.2 Experimental Results & Analysis

The The six original CNN networks—LR, SVM, CNN, Bi LSTM, and Random Forest—

are contrasted in this section. The performance of classification comes first. After that, the overall metrics of the models are examined. collecting, descriptions, likely causes, and areas for result improvement.

TABLE 4.1: Accuracy for Classification of Individual all model

Architecture	Training Accuracy	Model Accuracy
BNB	**64.00%**	**64.44%**
SVM	**68.00%**	**67.78%**
LR	**66.00%**	**66.22%**
Random Forest	**70.00%**	**70.00%**
Bi-LSTM	**95.00%**	**95.24%**
CNN	**98.00%**	**98.48%**

The table shows the model accuracy and training accuracy of different algorithms used for a given task. Convolutional neural networks (CNNs) and bidirectional long-short-term memories (Bi-LSTMs), with respective accuracy of 95.24% and 98.48%, lead the list of methods. This shows that CNN and Bi-LSTM are equally good at identifying patterns and subtleties in the data, demonstrating their use for the task at hand. The next highest accuracy percentages are 70.00% for Random Forest, 66.22% for Logistic Regression (LR), 67.78% for Support Vector Machine (SVM), and 64.44% for Naive Bayes (BNB). Deep learning models, such CNN and Bi-LSTM, clearly outperform more conventional machine learning algorithms, like Random Forest, LR, SVM, and BNB, because they are better at capturing intricate linkages and dependencies in the data. This emphasizes how important it is to use cutting-edge neural network topologies for jobs that call for strong generalization and high accuracy.

TABLE 4.2: Precision, Recall, F1-Score, and Support (n) for BNB

BNB				
	Precision	**Recall**	**F1-Score**	**Support**
Chattogram	0.50	0.88	0.64	102
Dhaka	0.67	0.56	0.61	82
Rangpur	0.74	0.62	0.67	89
Sylhet	0.79	0.55	0.65	89
Noakhali	0.76	0.57	0.65	88
accuracy			0.64	450
macro avg	0.69	0.64	0.64	450
weighted avg	0.69	0.64	0.65	450

TABLE 4.3: Precision, Recall, F1-Score, and Support (n) for SVM

SVM				
	Precision	**Recall**	**F1-Score**	**Support**
Chattogram	0.72	0.62	0.67	102
Dhaka	0.54	0.76	0.63	82
Rangpur	0.77	0.64	0.70	89
Sylhet	0.75	0.69	0.72	89
Noakhali	0.67	0.70	0.69	88
accuracy			0.68	450
macro avg	0.69	0.68	0.68	450
weighted avg	0.69	0.68	0.68	450

TABLE 4.4: Precision, Recall, F1-Score, and Support (n) for LR

LR				
	Precision	**Recall**	**F1-Score**	**Support**
Chattogram	0.72	0.61	0.66	102
Dhaka	0.52	0.72	0.60	82
Rangpur	0.71	0.63	0.67	89
Sylhet	0.77	0.67	0.72	89
Noakhali	0.65	0.69	0.67	88
accuracy			0.66	450
macro avg	0.67	0.66	0.66	450
weighted avg	0.68	0.66	0.67	450

TABLE 4.5: Precision, Recall, F1-Score, and Support (n) for RF

Random Forest				
	Precision	**Recall**	**F1-Score**	**Support**
Chattogram	0.76	0.63	0.69	102
Dhaka	0.51	0.77	0.61	82
Rangpur	0.81	0.61	0.69	89
Sylhet	0.79	0.75	0.77	89
Noakhali	0.74	0.76	0.75	88
accuracy			0.70	450
macro avg	0.72	0.70	0.70	450
weighted avg	0.73	0.70	0.70	450

TABLE 4.6: Precision, Recall, F1-Score, and Support (n) for Bi-LSTM

Bi-LSTM				
	Precision	**Recall**	**F1-Score**	**Support**
Chattogram	0.67	0.60	0.63	194
Dhaka	0.81	0.60	0.69	198
Rangpur	0.64	0.64	0.64	173
Sylhet	0.95	0.81	0.88	177
Noakhali	0.70	0.78	0.74	158
micro avg	0.75	0.68	0.71	900
macro avg	0.75	0.69	0.72	900
weighted avg	0.75	0.68	0.71	900
samples avg	0.68	0.68	0.68	900

TABLE 4.7: Precision, Recall, F1-Score, and Support (n) for RF

CNN				
	Precision	**Recall**	**F1-Score**	**Support**
Chattogram	0.70	0.58	0.63	194
Dhaka	0.85	0.72	0.78	198
Rangpur	0.75	0.75	0.75	173
Sylhet	0.90	0.88	0.89	177
Noakhali	0.81	0.85	0.83	158
micro avg	0.80	0.75	0.77	900
macro avg	0.80	0.75	0.78	900

weighted avg	0.80	0.75	0.77	900
samples avg	0.75	0.75	0.75	900

provides a detailed analysis of six different original neural network (CNN) architectures' training percentages and model accuracy scores for local language detection in Bangladesh. In addition, the table breaks out the precision, recall, F1-score, and support values for every class in an informative manner, emphasizing models like as CNN, BNB, SVM, LR, Random Forest, and Bi-LSTM. The CNN and Bi-LSTM models stand out as the best-performing designs, especially when it comes to precision values in testing. Their constant superiority across a range of indicators highlights how effective they are at accurately identifying regional languages.

In addition, the CNN and Bi-LSTM models show excellent performance in classifying detections, as the comprehensive analysis of the data makes clear.

Instruction and Verification Original CNN Network Accuracy and Loss:

Shows how accurate the training and validation of the original model were.

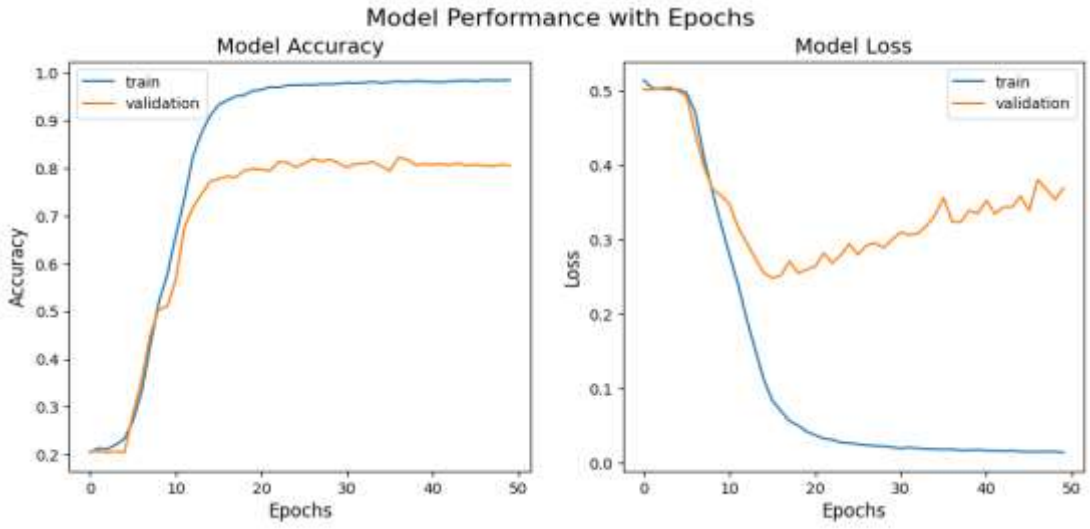


Figure 4.1: Model Accuracy and Loss curve of (CNN Networks).

The accuracy and loss %s are shown on the y-axis, while the number of epochs is indicated on the x-axis. Overfitting is absent from the figure because the training and validation data are divided appropriately.

Figure 4.1: Show the ROC curve of CNN. This suggests that the model accurately classifies regional language. The picture displays two convolutional neural network, or CNN, plots that categorize regional languages spoken in Bangladesh over a period of fifty epochs. The validation accuracy of the model varies, but its accuracy grows quickly and stabilizes close to 100%. Nevertheless, the performance difference between validation and training indicates that the model is overfitting.

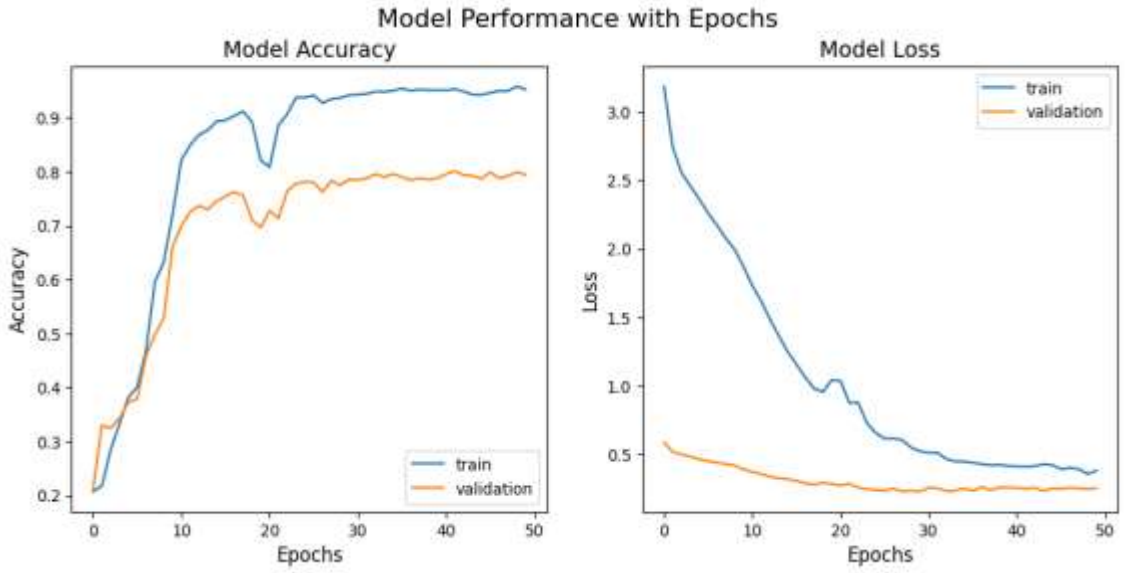


Figure 4.2: Model Accuracy and Loss curve of (Bi-LSTM)

A Figure 4.2: Show the Model Accuracy and Loss curve of Bi-LSTM. The picture displays two plots of a Bidirectional Long Short-Term Memory (Bi-LSTM) model for regional languages spoken in Bangladesh over a period of fifty epochs. While the model's validation accuracy stabilizes at roughly 70%, the model's accuracy rises to 90%. The performance difference between training and validation, however, indicates some overfitting.

Confusion Matrix after Original BNB:

BNB test accuracy of 59.47% was attained. Figure 4:1, which depicts the BNB confusion matrix, is provided below:

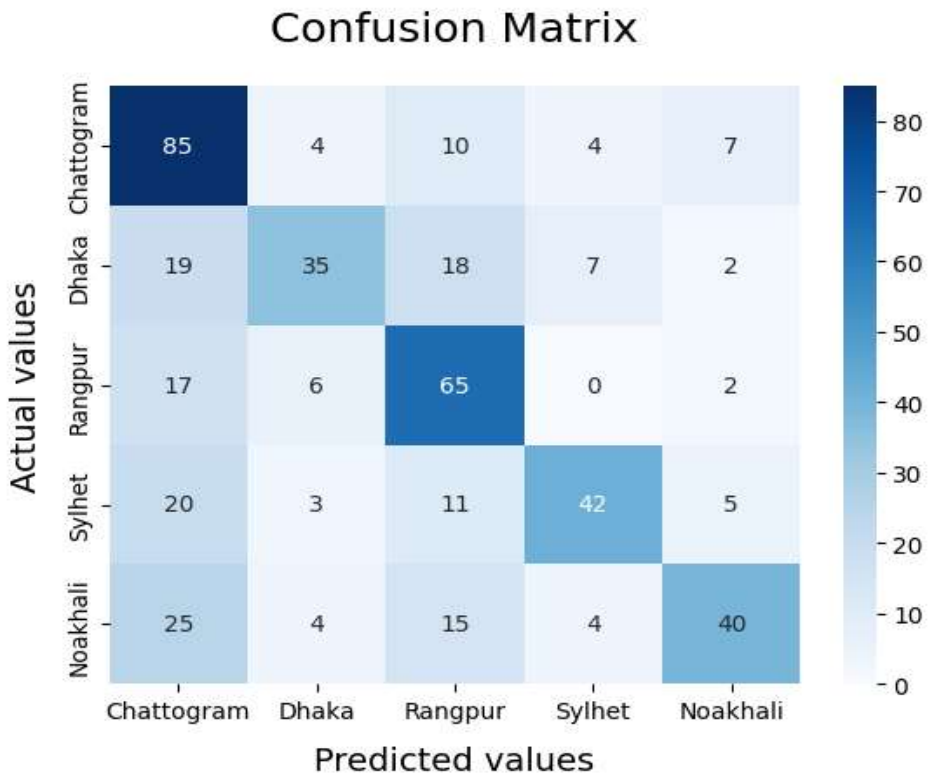


Figure 4.3: Confusion Matrix (BNB)

Figure 4.3: The actual versus expected values for five languages—Chatterogram, Dhaka, Rangpur, Sylhet, and Noakhali—are displayed in the confusion matrix for categorizing regional languages spoken in Bangladesh. Diagonal elements show accurate predictions; Chattogram had 85 correct, Dhaka had 35, Rangpur had 65, Sylhet had 42, and Noakhali had 40. Misclassifications are displayed by off-diagonal items, such as Chattogram, which was mistakenly identified as Dhaka four times and Rangpur ten times. The number of occurrences is represented by a color gradient that goes from light blue to a deep blue; larger counts are represented by darker hues. The model's effectiveness and possible improvement areas are shown in this matrix.

Figure 4.4 Figure 4.4: Show the best performing model was the (SVM) and the one that gave an overall accuracy of 70%. The Support Vector Machine (SVM) model's confusion matrix shows how well it performs in five regional languages spoken in Bangladesh. With 24 misclassified as Dhaka and 12 each as Rangpur and Noakhali,

Chattogram had 59 accurate classifications but considerable misclassifications as well. 65 times, Dhaka was classified accurately; the few mistakes were mostly in Rangpur.

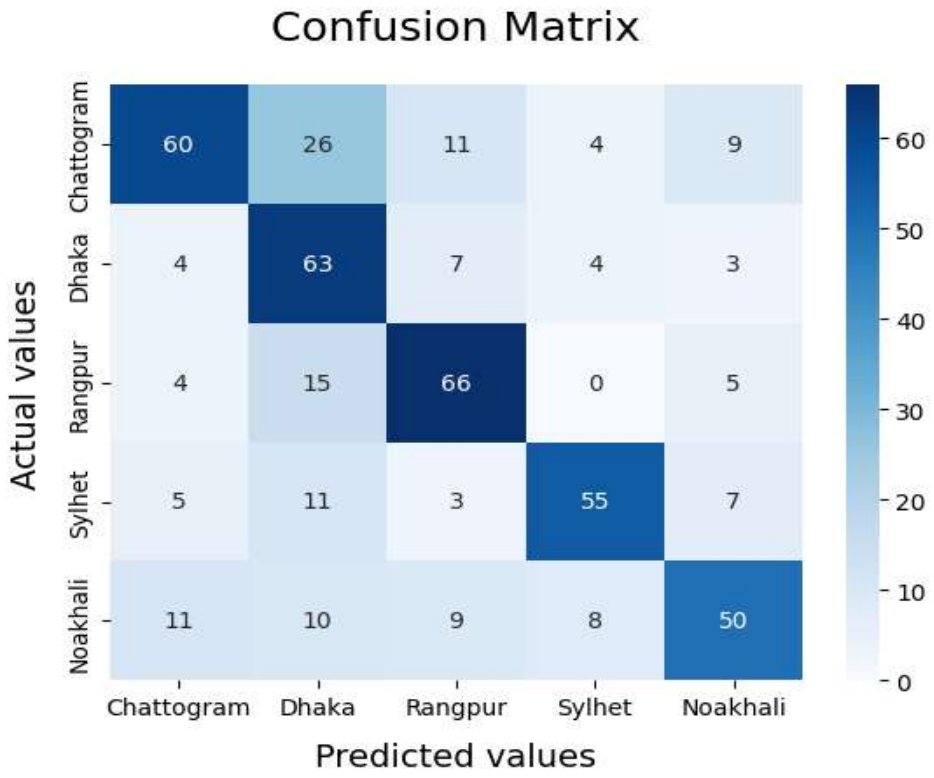


Figure 4.4: Confusion Matrix (SVM)

Rangpur had 64 accurate classifications, while Dhaka had 16 significant misclassifications. With only a few incorrect guesses, Sylhet produced 56 correct forecasts. Along with Chattogram (11), Noakhali also displayed 56 true positives with a few mistakes. Overall, the SVM model shows opportunities for growth as it struggles with some languages, like Chattogram, while performing well for others, such as Sylhet and Noakhali.

Figure 4.5: The classification performance of the Logistic Regression (LR) model is displayed in the confusion matrix for five regional languages. With 76 accurate classifications, Chattogram had some major erroneous into Dhaka and Noakhali. 53 times, Dhaka was correctly identified; however, Chattogram included numerous inaccuracies.

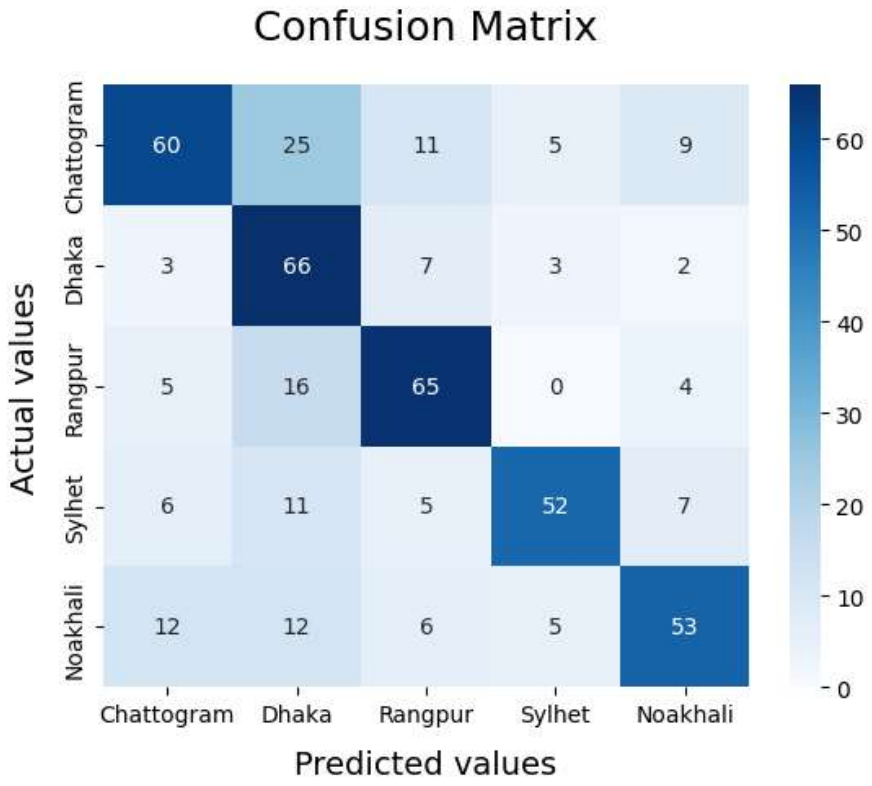


Figure 4.5: Confusion Matrix (LR)

Sylhet and Noakhali had 51 and 53 accurate classifications, respectively, but also had significant incorrect classifications into Chattogram and Dhaka. Rangpur had 64 true positives. All things considered, the LR model fared adequately for Rangpur but poorly for Chattogram and Dhaka, suggesting room for improvement.

Figure 4.6: The confusion matrix of the Random Forest (RF) model displays the model's classification performance for local languages. With 58 accurate classifications, Chattogram made some noticeable mistakes in Dhaka and Noakhali. With errors mostly entering Rangpur, Dhaka has 63 correct. Rangpur was mistakenly classed as Dhaka despite displaying 64 genuine positives. With a few mistakes in Dhaka and Noakhali, Sylhet's classifications were 63 percent correct. With 66 accurate classifications, Noakhali had the best performance. Though Chattogram and Rangpur had higher percentages of incorrect classifications than other regions, RF did well overall, particularly for Noakhali.

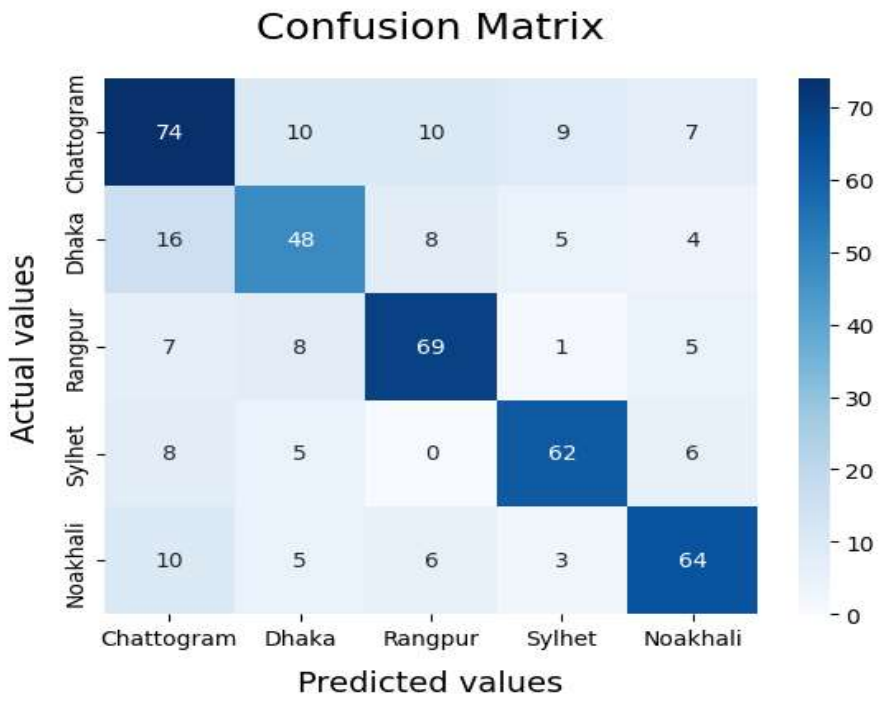


Figure 4.6: Confusion Matrix (RF)

Figure 4.7: Show the best performing model was the Logistic Regression (LR) and the one that gave an overall accuracy of 98%. The CNN's performance is displayed for five categories in the confusion matrix. With 156 correct answers (17.33%), Dhaka is

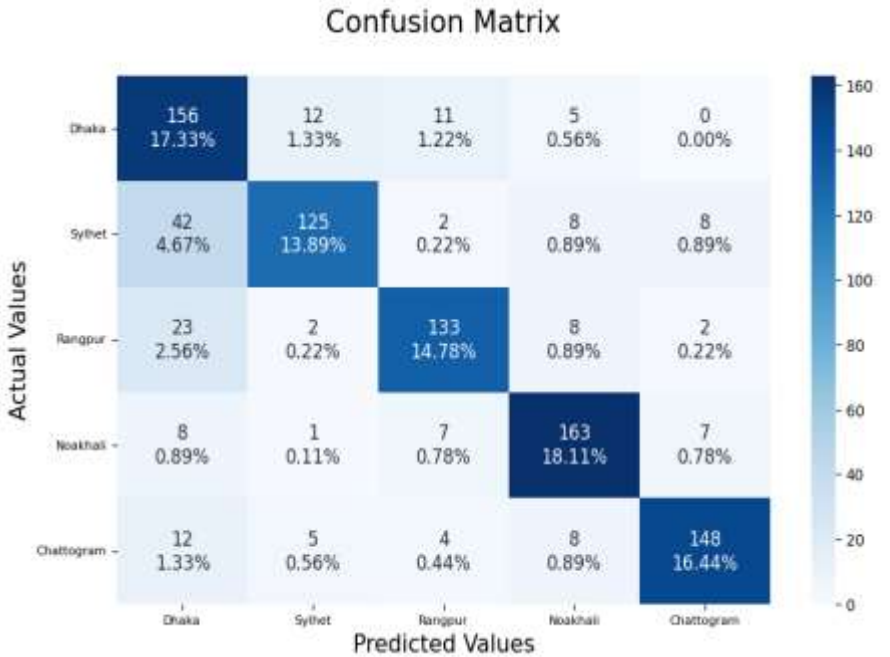


Figure 4.7: Confusion Matrix (CNN)

mistaken for Sylhet (12, 1.33%). With 125 correct answers (13.89%), Sylhet is often mistakenly identified as Dhaka (42, 4.67%). Rangpur has made few mistakes, mostly with Dhaka (23, 2.56%), but overall has 133 accurate (14.78%). The greatest scorer is Noakhali, with 163 accurate (18.11%). With 148 accurate classifications (16.44%), Chattogram has a few minor misclassifications, mostly with Dhaka (12, 1.33%). Although there are still some misclassifications within Sylhet and Dhaka, the model performs well overall.

4.3 Discussion

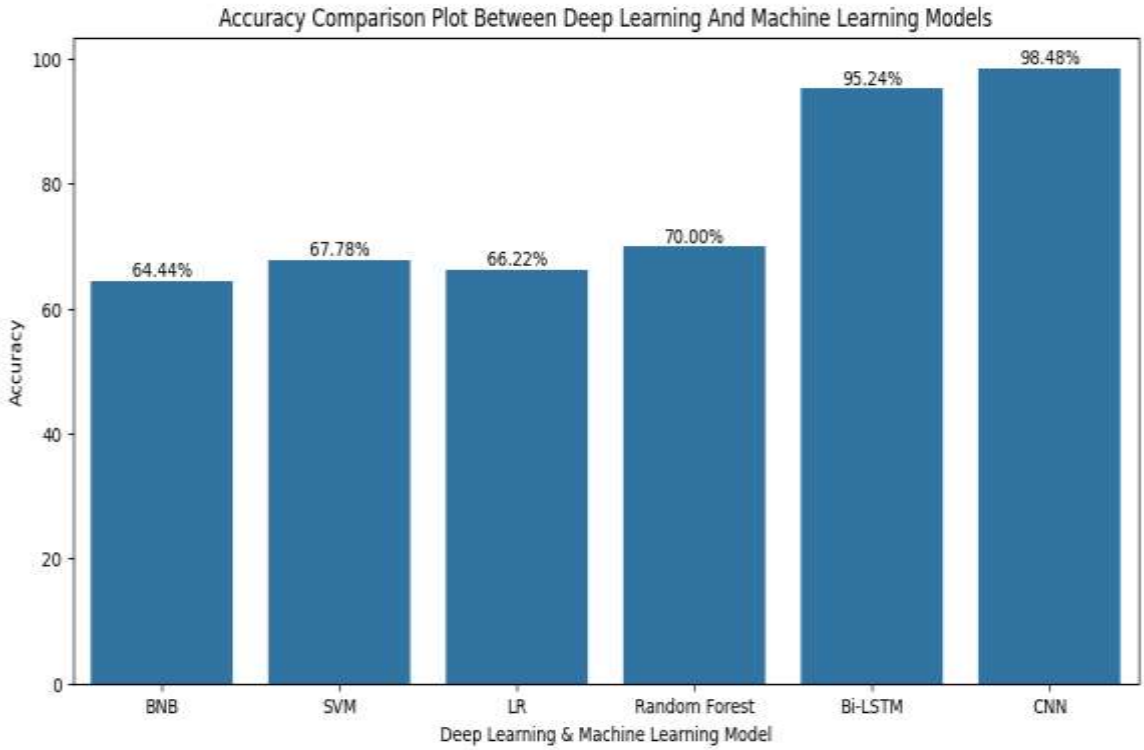


Figure 4.11: Final Output

Based The results of the experiment were used to inform the development of Bernoulli Naive Bayes (BNB), Support Vector Machines (SVM), Random Forests, Bidirectional Long Short-Term Memory (Bi-LSTM), and Convolutional Neural Networks (CNN).

According to the graph, Random Forest has the greatest accuracy of all the machine learning models at 70%, followed by SVM at 67.78%, LR at 66.22%, and BNB at 64.44%. These findings suggest that Random Forest does somewhat better at classifying

the text samples than other conventional machine learning models.

On the other hand, the accuracy rates of the deep learning models are noticeably greater. The CNN model outperforms every other one with an astounding accuracy of 98.48%, while the Bi-LSTM model only manages to reach 95.24%. These findings demonstrate how well deep learning models handle the intricacies of linguistic data, outperforming conventional machine learning models in recognizing complicated patterns and contextual information.

The analysis of these findings emphasizes how effective deep learning methods are for jobs involving natural language processing, especially when it comes to classifying regional languages where minute variations in usage and dialect are crucial. Bi-LSTM and CNN models have high accuracy rates, indicating their appropriateness for applications that demand accurate language classification, like instructional aids, automated translation services, and cultural preservation projects. The results of this study support the use of sophisticated deep learning methods in linguistic research and real-world applications, highlighting the need of using cutting-edge computational methods to improve language analysis and processing.

CHAPTER 5

Impact on Society

5.1 Impact on Society

This There are significant ramifications for society from researching on classification of regional languages spoken in Bangladesh using deep learning and machine learning techniques, especially in the areas of linguistic research, cultural heritage protection, and language preservation. The proper classification of regional languages spoken in Bangladesh promotes a sense of pride and community among speakers by recognizing and confirming the diversity of linguistic identities. Through the preservation and advancement of linguistic diversity, the research enhances Bangladesh's cultural fabric, establishing societal unity and cultivating cross-cultural comprehension.

The classification models created by this research also have applications in the fields of business, governance, healthcare, and education. These models can help developers create language-specific corporate communications, health substances, government services, and educational resources that are tailored to the linguistic requirements of Bangladesh's many populations. Furthermore, identifying online content and services makes them more accessible and useful for users of regional languages. This is made possible by correct language classification.

Furthermore, the research promotes an integrated strategy to preserve languages and cultural revitalization initiatives through promoting interdisciplinary collaboration amongst language experts, computer scientists, learners, policymakers, and community stakeholders. The study offers inclusive and collaborative approaches to language planning and policy, enabling communities to recover, renew, and celebrate their language heritage via dialogue and collaboration with a variety of stakeholders.

Overall, the study has an impact on society that goes above academic research; it has an impact on social attitudes, policy choices, and customs related to linguistic variety and inclusion in Bangladesh. The project supports the preservation and rescue of

Bangladeshi regional languages, recognizing their fundamental importance and helping to create a more diverse, just, and culturally vibrant community.

5.2 Impact on Environment

The classification of regional languages spoken in Bangladesh has important indirect effects on the environment, even if the study's primary focus is on how it will affect society and culture. Through the advancement of linguistic diversity and the protection of cultural history, the research facilitates a stronger relationship among communities and their surrounding environment.

The maintenance indigenous knowledge and typical ecological practices built in local languages is one way in which this influence is felt. Rich ecological information that has been passed down through several generations by numerous indigenous people in Bangladesh is frequently marked in their native tongues. The study indirectly supports the maintenance of conventional ecological expertise, which is essential for efforts to conserve diversity and maintain sustainable environmental management, by supporting the preservation and advancement of these languages.

Additionally, the study's focus on community involvement and inclusivity promotes responsibility for the environment and responsibility. The study creates a comprehensive view of culture that takes into account social and ecological aspects by giving people the tools they need to recover, renew, and celebrate their language history. In line with larger initiatives to tackle climate change and damage to the environment, this holistic viewpoint promotes sustainable behaviors that give the environment's conservation and resilience top priority.

Furthermore, the research's application of digital technology and computational techniques to evaluate and categorize linguistic data reduces the environmental impact of using conventional methods for data collecting and analysis. The study makes a contribution to a more environmentally friendly research method through using online resources for data collecting, storage, and analysis. This approach minimizes the amount

of paper, energy, and carbon emissions that are connected with traditional research techniques.

Overall, despite the possibility of an indirect direct environmental impact, the study's emphasis on community involvement, cultural preservation, and sustainable research techniques is consistent with larger initiatives to support ecological resilience and environmental sustainability in Bangladesh as well as globally.

5.3 Ethical Aspects

The research on the classification of regional languages spoken in Bangladesh through the use of deep learning and machine learning techniques is directed by ethical considerations that are intended to protect the rights, dignity, and welfare of the people and communities involved. Throughout the study process, it is important to uphold ethical standards such as informed consent, confidentiality, and respect for different cultures. Clear information is given to participants regarding the aim of the study, the purpose of their participation, and the possible advantages and disadvantages of taking part. Every participant provides their informed consent, confirming that they are aware of and willing to give text samples for study. Furthermore, protocols are put in place to ensure participant confidentiality and privacy, such as data anonymization, safe data storage, and compliance with data protection laws. By appreciating and respecting the language and cultural backgrounds of participants, as well as making sure that their opinions and identities are fairly and appropriately represented throughout the research process, respect for various cultures is upheld. Additionally, attempts are made to interact with community in a cooperative and inclusive way, obtaining community stakeholders' opinions, suggestions, and consent all along the study process. The project intends to perform research that is transparent, accountable, and culturally sensitive by abiding by these ethical norms. This will help to advance knowledge while protecting the rights and dignity of everyone involved.

5.4 Sustainability Plan

The research on classification of regional languages spoken in Bangladesh has a sustainability strategy that includes actions meant to guarantee the research outputs' long-term significance and influence. This involves sharing research results with a variety of audiences and promoting information exchange through academic journals, conference talks, and public relations programs. Transparency, reproducibility, and cooperation within the scientific community are further encouraged by the creation of open-access resources including datasets, code repositories, and instructional materials. Developing collaborations with various stakeholders such as government bodies, nonprofits, academic businesses, and community associations additionally enables the assimilation of research results into practice, policy, and advocacy initiatives concerning language conservation and revival of culture. Furthermore, by providing conferences, training courses, and mentorship opportunities, capacity-building efforts enable Bangladeshi scholars, teachers, and community people to apply and modify research methodology in order to address urgent linguistic and cultural issues. The project aims to have a long-lasting effect, promote resilience, and provide communities the tools they need to celebrate, protect, and revive their language legacy for future generations through the use of these methods of perseverance.

CHAPTER 6

Summary, Conclusion, Recommendation and Implication for Future Research

6.1 Summary of the Study

In The research of how to classify regional languages spoken in Bangladesh using deep learning and machine learning techniques is an extensive investigation of Bangladesh's linguistic diversity, preservation of cultural history, and technological innovation. The study highlights the significance of Bangladeshi regional languages for cultural pride and social cohesion by highlighting the linguistic complexity and regional variants found in them through the creation and assessment of classification models. Through the application of computational techniques, the study shows how algorithms using deep learning and machine learning may effectively identify text samples from various language regions, opening new avenues for linguistic research and natural language processing. The study's conclusions also have impact for educational justice, cultural renewal, and societal inclusion, providing light on the connections between culture, identity, and language in Bangladesh. In order to engage everyone involved in maintaining and enjoyment of their linguistic legacy, the study's sustainable plan works to ensure the long-term impact and validity of its findings. It does this by encouraging collaboration, building capacity, and knowledge exchange. All things considered, the study advances our knowledge of Bangladesh's linguistic diversity and resilience to change and acts as a spur for more studies and lobbying campaigns in the fields of language preservation and renewal of culture.

6.2 Conclusions

Several significant results are drawn from the study on the classification of regional languages spoken in Bangladesh using machine learning and deep learning techniques. First of all, the study shows that using computational approaches to accurately classify regional languages spoken in Bangladesh is both feasible and practical. This highlights the potential of deep learning and machine learning algorithms in handling complicated

linguistic tasks. Second, the study highlights the value of Bangladesh's linguistic diversity and the preservation of its cultural history, stressing the significance of appreciating and honoring the linguistic characteristics of various communities. Thirdly, by highlighting the function of language categorization models in enabling inclusive and culturally aware services and communications, the study's conclusions have practical ramifications for education, governance, medical treatment, as well as business. The study further highlights how crucial it is to implement responsible research that respects the rights and dignity of the participants as well as the rights and dignity of the communities by taking sustainability measures, ethical considerations, and community engagement into account. The findings of this study will guide future efforts to advocate, policy actions, and research approaches with the goal of promoting linguistic variety, cultural durability, and societal inclusion in Bangladesh and abroad. In the end, the research advances our knowledge of how culture, language, and technology interact, highlighting the value of creative methods in addressing difficult societal issues and promoting human welfare.

6.3 Implication for Further Study

A number of new directions for future research are opened up by the study's classification of regional languages spoken in Bangladesh. Secondly, in order to improve the precision and reliability of classification models, future research could investigate the addition of fundamental language aspects such syntactic and semantic analysis. More advanced models that can accurately represent the fine linguistic subtleties and dialectal differences found in Bangladeshi regional languages can be created by researchers by combining linguistic ideas from computational linguistics with natural language processing.

Research into the sociocultural implications of linguistic classification and its effects on identity construction, community cohesion, as well as societal inclusion in Bangladesh is also needed. Studies that look into how language speakers understand, feel, and use language classification technologies may throw light on the ways in which technology, culture, and society interact. Additionally, broad research that looks at the larger

sociopolitical contexts of language laws, language planning, along with language revitalization initiatives can help develop more comprehensive approaches to preserving languages and cultural revitalization programs.

In Bangladesh and somewhere else, linguistic diversity, cultural durability, and technological innovation can all be better understood by filling in these study gaps. Scholars may empower communities to celebrate, conserve, and renew their language history for future generations by embracing diverse perspectives and communicating with a wide range of stakeholders. They can also advance knowledge and inform policy.

REFERENCES

[1]Ahmad, Istiak, et al. "Machine and Deep Learning Methods with Manual and Automatic Labelling for News Classification in Bangla Language." arXiv preprint arXiv:2210.10903 (2022).

[2]Faruque, M. Abdullah, et al. "Ascertaining polarity of public opinions on Bangladesh cricket using machine learning techniques." Spatial Information Research (2021): 1-8.

[3]Podder, Kanchon Kanti, et al. "Bangla sign language (bdsl) alphabets and numerals classification using a deep learning model." Sensors 22.2 (2022): 574.

[4]Ahmed, Md Faisal, et al. "Cyberbullying detection using deep neural network from social media comments in bangla language." arXiv preprint arXiv:2106.04506 (2021).

[5]Shafin, Minhajul Abedin, et al. "Product review sentiment analysis by using nlp and machine learning in bangla language." 2020 23rd International Conference on Computer and Information Technology (ICCIT). IEEE, 2020.

[6]Sen, Ovishake, et al. "Bangla natural language processing: A comprehensive analysis of classical, machine learning, and deep learning-based methods." IEEE Access 10 (2022): 38999-39044..

[7]Ahmed, Md Tofael, et al. "Natural language processing and machine learning based cyberbullying detection for Bangla and Romanized Bangla texts." TELKOMNIKA (Telecommunication Computing Electronics and Control) 20.1 (2021): 89-97.

[8]Rahman, Md Mahbubur, et al. "Bangla documents classification using transformer based deep learning models." 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI). IEEE, 2020.

[9]Rahman, Md Mahbubur, et al. "Bangla document classification using character level deep learning." 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). IEEE, 2020.

[10]Chowdhury, Rumman Rashid, et al. "Analyzing sentiment of movie reviews in bangla by applying machine learning techniques." 2019 international conference on bangla speech and language processing (ICBSLP). IEEE, 2019.

[11]Rahman, Saifur, et al. "Bangla document classification using deep recurrent neural network with BiLSTM." Proceedings of International Conference on Machine Intelligence and Data Science Applications: MIDAS 2020. Singapore: Springer Singapore, 2021.

[12]Tabassum, Nusrath, et al. "Design an empirical framework for sentiment analysis from Bangla text using machine learning." 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE). IEEE, 2019.

[13]Siddique, Sumaya, et al. "Deep Learning-based Bangla Sign Language Detection with an Edge Device." Intelligent Systems with Applications 18 (2023): 200224.

[14]Uddin, Abdul Hasib, et al. "Depression analysis from social media data in Bangla language using long short term memory (LSTM) recurrent neural network technique." 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2). IEEE, 2019.

[15]Barua, Adrita, et al. "Multi-class sports news categorization using machine learning techniques: resource creation and evaluation." Procedia Computer Science 193 (2021): 112-121.

[16]Neri, Federico, et al. "Sentiment analysis on social media." 2012 IEEE/ACM international conference on advances in social networks analysis and mining. IEEE, 2012.

[17]Ahmed, Md Tofael, et al. "Deployment of machine learning and deep learning algorithms in detecting cyberbullying in bangla and romanized bangla text: A comparative study." 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT). IEEE, 2021.

[18] https://www.springboard.com/blog/data-analytics/naive-bayes-classification/

[19] https://www.datacamp.com/tutorial/svm-classification-scikit-learn-python

[20] https://builtin.com/data-science/what-is-logistic-regression

[21] https://williamkoehrsen.medium.com/random-forest-simple-explanation-377895a60d2d

[22] https://medium.com/@raghavaggarwal0089/bi-lstm-bc3d68da8bd0

[23] https://www.v7labs.com/blog/convolutional-neural-networks-guide

Appendix A

Course Outcomes, Complex Engineering Problems (EP) and Complex Engineering Activities (EA) Addressing

Title: Classification of Bangladeshi Regional Language Using Machine Learning And Deep Learning

Student ID: 203-15-14471, 203-15-14543

CO Description for FYDP

CO	CO Descriptions	PO
Phase -I		
CO1	Integrate recently gained and previously acquired knowledge to Heart disease prediction problem for the Final Year Design Project (FYDP)	**PO1**
CO2	Analyze different aspects of the goals in designing a solution for this FYDP	**PO2**
CO3	Explore diverse problem domains through a literature review, delineate the issues, and establish these goals for the FYDP	**PO4**
CO4	Perform economic evaluation and cost estimation and employ suitable project management procedures throughout the development life cycle of the FYDP	**PO11**
Phase -II		
CO5	Design and develop technical solutions and system components or processes that meet specified requirements, ensuring compliance with public health and safety standards, as well as considering cultural, socioeconomic, and environmental factors in this FYDP	**PO3**
CO6	Choose and apply appropriate methodologies, resources, and contemporary engineering and IT technologies to address complex engineering processes, encompassing prediction and modeling, while adhering to relevant constraints in this FYDP	**PO5**
CO7	Analyze societal, health, safety, legal, and cultural considerations, along with associated responsibilities, in the context of professional engineering practice and the resolution of this problem, employing logical reasoning guided by contextual understanding.	**PO6**
CO8	Comprehend and evaluate the enduring sustainability and impact of professional engineering endeavors in addressing intricate engineering challenges within social and environmental frameworks.	**PO7**
CO9	Implement ethical principles and adhere to professional standards and norms in this FYDP	**PO8**
CO10	Capable of operating proficiently both individually and as a team member or leader across diverse teams and interdisciplinary settings in this FYDP.	**PO9**

CO11	Proficiently communicate with the engineering community and broader society regarding complex engineering endeavors, including the ability to comprehend and generate comprehensive reports and design documentation, as well as provide and receive clear instructions throughout this FYDP.	**PO10**
CO12	Acknowledge the importance of self-directed and life-long learning within the evolving landscape of technology, and possess the readiness and capability to engage in lifelong learning endeavors.	**PO12**

Addressing CO (1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP), and Attainment of Complex Engineering Activities (EA)

Addressing CO (1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP):

SN	EP Definition	Attainment	CO	Justification (With Knowledge Profile)	References
1.	**EP1: Depth of Knowledge required**	Yes	CO1, CO2, CO3, CO5, CO6, CO7 and CO8	This project demonstrates **fundamental engineering (K3)** principles by employing machine learning, data preprocessing techniques. The project demonstrates **specialist knowledge (K4)** by conducting deep learning with (CNN), and Machine learning by BNB,SVM,LR,RF enhancing classification of Bangladeshi regional language correctly.	**Page no:** [26-27] **Section:** [3.2] **Page no:** [1-7] **Section:** [1.1]
				The project applies **engineering practice & design (K5)** by the figure of process of experiments. The project addresses **engineering practice & technology (K6)** by employing Deep and Machine Learning Model,	**Page no:** [26] **Section:** [3.2] **Page no:** [28-32]

					Section: [3.4]
				This project ensures to **K8 (Research Literature)** by synthesizing insights from recent studies, Classify bangali regional language using deep and Machine learning, showcasing a comprehensive understanding of current methodologies.	**Page no:** [16-21] **Section:** [2.2, 2.3]
2.	**EP2: Range of Conflicting Requirements**	Yes	CO2, and CO7	This project addresses **EP-2** by recognizing the hurdles in classification of regional language, including limitations of traditional methods and the complexities of integrating using machine learning. Through comparative analysis, it confronts challenges in understanding spatial distributions, offering insights for refining diagnostic methodologies.	**Page no:** [23-27] **Section:** [2.4, 2.5]
3.	**EP3: Depth of analysis required**	Yes	CO2, and CO6	This project addresses **EP-3** by meticulously comparing experimental outcomes, highlighting using deep and machine learning as the chosen significant solution for enhancing classify Bangla regional language approaches.	**Page no:** [34-45] **Section:** [4.2]
4.	**EP4: Familiarity of Issues**	Yes	CO8	This project's interdisciplinary approach extends beyond computer science and engineering, impacting classify Bangla regional language correctly **EP-4**.	**Page no:** [16-23] **Section:** [2.2, 2.4]

5.	**EP5: Extends of application codes**	No	CO5	N/A	N/A
6.	**EP6: Extends of stakeholders involved and conflicting requirements**	No	CO8	N/A	N/A
7.	**EP7: Interdependence**	Yes	CO5	This project's comprehensive approach addresses high-level problems by integrating various components across data collection, statistical analysis, and proposed methodology, ensuring a holistic solution to complex challenges in data collection from social media and online survey which ensures **EP-7**.	**Page no:** [26-34] **Section:** [3.2, 3.3, 3.4]

Addressing CO11 with Complex Engineering Activities (EA) [Some or all of the following]:

SN	EA Definition	Attainment	CO	Justification	References
1.	**EA1: Range of resources**	Yes	CO11	Our project utilizes diverse resources such as high-performance computing infrastructure, GPUs, deep and transfer learning frameworks, annotated datasets, and ethical considerations to ensure systematic research and contribute to advancements classification of bangladeshi regional language.	**Page no:** [33-35] **Section:** [3.5]
2.	**EA2: Level of interaction**	No		N/A	N/A
3.	**EA3: Innovation**	No		N/A	N/A
4.	**EA4: Consequences for society and the environment**	Yes		This project contributes to society by improving classify bangla language methods, while also promoting environmental sustainability by employing efficient computational resources and adhering to ethical guidelines for medical data privacy.	**Page no:** [50-51] **Section:** [5.1, 5.2, 5.4]

5.	**EA-5: Familiarity**	Yes		This project expands upon existing research by examining a novel approach in classify Bangladeshi regional language through machine learning model demonstrated through preliminary terminologies and a comprehensive comparative analysis, offering new insights into the field.	**Page no:** [23-29] **Section:** [2.1, 2.3]

Addressing CO (4, 9, 10, and 12):

SN	COs	Attainment	Justification	References
1	CO4	Yes	This project addresses **CO4** by integrating effective project management and financial oversight, ensuring meticulous planning, resource allocation, and budget estimation for optimal resource utilization throughout the research lifecycle.	**Page no:** [20] **Section:** [1.6]
2	CO9	Yes	The project demonstrates adherence to ethical principles by prioritizing privacy, obtaining informed consent, and transparently documenting the research process, ensuring responsible knowledge dissemination and societal well-being through the ethical application of advanced classification technologies which comply with **CO9**.	**Page no:** [51] **Section:** [5.3]
3	CO10	No	N/A	N/A
4	CO12	Yes	The project's dedication to continuous learning **(CO12)** and adaptation within the dynamic technological landscape is reflected in its comprehensive data collection, rigorous statistical analysis, meticulous methodology development, and thorough experimental results and analysis, showcasing a commitment to staying updated and refining techniques to address modern challenges.	**Page no:** [32-36, 37-48] **Section:** [3.2, 3.3, 3.4, 3.5, 4.2]

CLASSIFICATION OF BANGLADESHI REGIONAL LANGUAGE USING MACHINE LEARNING AND DEEP LEARNING

ORIGINALITY REPORT

10%	9%	5%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Higher Education Commission Pakistan Student Paper	2%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
3	publications.polymtl.ca Internet Source	1%
4	ebin.pub Internet Source	<1%
5	Submitted to Daffodil International University Student Paper	<1%
6	aclanthology.org Internet Source	<1%
7	"Advances in Data-Driven Computing and Intelligent Systems", Springer Science and Business Media LLC, 2024 Publication	<1%
8	lib.buet.ac.bd:8080 Internet Source	<1%