Independent University **Bangladesh (IUB)**

IUB Academic Repository

Computer Science and Engineering | Undergraduate Thesis

2025-12-14

Multilingual sign language translation using image-based classification models for Bangla, American, and Indian languages with ResNet and EfficientNet

Biswas, Sudipto

Independent University, Bangladesh (IUB)

https://ar.iub.edu.bd/handle/11348/1579

Downloaded from IUB Academic Repository



Multilingual Sign Language Translation using image-Based classification models for Bangla, American, Indian Languages with ResNet and EfficientNet

By

Sudipto Biswas*

Student ID: 2222391*

Autumn, 2025

Supervisor:

Saadia Binte Alam, PhD

Associate Professor and Head

Department of Computer Science & Engineering

Independent University, Bangladesh

December 14, 2025

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

Department of Computer Science & Engineering

Independent University, Bangladesh

Attestation

We are conscious of the nature of plagiarism and the university's stringent antiplagiarism policy. We confirm that this is our original work. Moreover, any software or textual work by others that is utilized in this project is acknowledged appropriately and in accordance with generally accepted academic standards.

Signature	Date
Sudipto Biswas	
Name	

Acknowledgement

I would like to express my sincere gratitude to Saadia Binte Alam, PhD, for her guidance, support, and valuable feedback throughout the development of this thesis. I am also thankful to the members of the evaluation committee for their helpful comments. My appreciation goes to all my teachers in the Department of Computer Science and Engineering for the knowledge and skills I gained during my undergraduate studies, which helped me complete this work. I am grateful to my family for their continuous support and to my friends for encouraging me during this project. Finally, I would like to thank everyone who contributed in any way to this thesis.

Finally, I am thankful to everyone who contributed directly or indirectly to this work. Each form of support, big or small, played a part in helping me complete this thesis successfully.

Letter of Transmittal

December 14, 2025
Saadia Binte Alam, PhD
Senior Project Supervisor & Assistant Professor
Department of Computer Science and Engineering,
Independent University, Bangladesh.

Subject: Submission of Senior Project Report on "Multilingual Sign Language Translation using Image-Based Classification Models for Bangla, American, Indian Languages with ResNet and EfficientNet.".

Dear Sir,

It is a great privilege to submit my senior project report titled "Multilingual Sign Language Translation using Image-Based Classification Models for Bangla, American, Indian Languages with ResNet and EfficientNet." This report reflects the work, learning, and effort I have invested throughout the project under your kind supervision. During the development of this project, I explored different deep learning models, worked with multilingual datasets, and gained hands-on experience in building an image-based classification system for sign language translation. This journey has strengthened my understanding of computer vision, model training, and real-world challenges in assistive communication technologies. I believe the report presents the project's objectives, methods, results, and findings with clarity and sincerity. I am sincerely grateful for your continuous guidance, patience, and valuable feedback. Your support has been instrumental in helping me complete this work. I hope the report meets your expectations and is considered satisfactory.

Thank you for your time, encouragement, and assistance throughout the semester.

Sincerely,
Sudipto Biswas
2222391

Evaluation Committee

Supervision Panel

.....
Supervisor Saadia Binte Alam, PhD

Internal Examiners

.....
Person A

.....
Person B

External Examiners

.....
Person C

Contents

Attestation i

Acknowledgement ii

Letter of Transmittal iii

Evaluation Committee iv

1 Introduction 3

1.1 Background 3

1.2 Problem Statement 4

1.3 Objectives 4

1.4 Scopes 5

2 Literature Review 6

2.1 Related Works 6

3 Dataset Description 10

3.1 Datasets 10

3.1.1 Dataset Overview 10

3.1.2 Dataset Collection 12

3.1.3 Dataset Preparation 13
3.1.4 Data Preprocessing 13
3.1.5 Data Balancing and Preparation 14

4 Methodology **15**

4.1 List of tools used 15
4.1.1 Hardware Specifications 15
4.1.2 Software Specifications 15
4.2 Methods 16
4.2.1 Data Loading and Sorting 16
4.2.2 Normalization 17
4.2.3 Splitting the Dataset 17
4.2.4 Deep Learning Models (ResNet & EfficientNet) 18
4.2.5 Sign Translation Pipeline 19
4.2.6 Evaluation Metrics 19
4.2.7 Generalization Strategy 20

5 Results and Analysis **21**

5.1 Model Performance Evaluation 22
5.2 Analysis of Prediction Accuracy 23
5.3 Comparison Between ResNet and EfficientNet 24

6 Social Impact **26**

6.1 Sustainability of the Work 26
6.2 Social and Environmental Impact Analysis 27

6.2.1 Social Impact 27

6.2.2 Environmental Impact 27

6.3 Ethical Considerations and Framework 29

7 Lesson Learned **30**

7.1 Challenges Faced 31

7.2 Solutions to the Challenges 32

8 Future Work & Conclusion **33**

8.1 Conclusion 33

8.2 Future Work 34

Bibliography **36**

List of Figures

3.1 Sample images from the three sign language datasets used in this study. Each row shows three signs with their corresponding alphabet meaning: BdSL (top row), ISL (second row), and ASL (bottom row). 11

4.1 Research workflow showing the main steps followed in this study. 16

5.1 Predictions from EfficientNet-B0 and ResNet-50 with confidence scores and ground-truth labels. 23

5.2 Classes with accuracy < 0.8 in BdSL and ISL for all models 25

5.3 Average macro AUROC for all models across BdSL, ASL and ISL 25

List of Tables

4.1 List of Hardware Specifications 15

4.2 Software Libraries and Tools Used in This Study 16

5.1 Macro-average metrics for both models on the test dataset 21

Abstract

Sign languages play a vital role in the life of Deaf and Hard-of-Hearing people, as this type of communication enables them to share the ideas as visual signs rather than verbal or written components. Nevertheless, even though communication gaps are eliminated when the signers communicate with individuals with knowledge of the sign language, there is still the existence of communication gaps in the interaction of the signers with those who do not speak the sign language, particularly in emergency cases, public services, and education. The mentioned challenges demonstrate the necessity of simpler and more affordable tools capable of identifying sign language and converting it into a different format that could be understood by other people. The proposed study is a multilingual image recognizing system to find out hand signs in three popular sign languages: Bangla Sign Language (BdSL), American Sign Language (ASL) and Indian Sign Language (ISL). The goal is to develop a useful and portable model that is able to categorize the static hand gestures and assist future translation applications. The two deep learning models, ResNet50 and EfficientNetB0, were chosen in this paper because of their accuracy, speed, and the cost of computing. Having gathered and processed the sets of data of the three sign languages, all the pictures were scaled, normalized, and split into training, validation, and testing subsets. The two models were trained in the same environment in order to come up with a fair comparison. ResNet50 was found to be highly effective in learning finer visual details, which is why it is efficient in differentiating the signs that have similar shapes. EfficientNetB0 was very competitive in accuracy and the computation cost was much lower hence it can be used in lightweight or portable systems. These findings have shown that it is possible to develop an image-based multilingual sign language recognition system with the help of the latest deep learning models. The study is associated with limitations despite the positive results it demonstrates. The system can only be trained on still images and is not able to understand continuous gestures, hand movements, and the expressions at the sentence level. The consistency of predictions in the real world can also be influenced by differences in lighting conditions, camera angles, and variations among datasets in the signers. Moreover, the available work is limited to three of the sign languages and also tends to recognize alphabets instead of words

or phrases. In order to overcome these shortcomings, the future research could employ video-based data that can record movement and enable models to acquire time-related information. To enhance the ability to recognize dynamic gestures, it may be advisable to integrate sequence-learning methods, including transformer-based architectures or LSTM-based networks. The framework can be made more practical and applicable in the daily communication by expanding the system to accommodate more sign languages and creating a real-time mobile or web-based application. The system can be further improved, including better data augmentation and signer-independent trainers, which can assist the system in generalizing more effectively to a wide range of users. Such initiatives are eventually likely to result in a multilingual sign language translation option that is more inclusive and hence allows easier communication between Deaf people and the rest of the society.

Keywords— Sign Language Recognition, Multilingual Translation, Bangla Sign Language (BdSL), American Sign Language (ASL), Indian Sign Language (ISL), Deep Learning, Image Classification, ResNet50, EfficientNetB0, Computer Vision.

Chapter 1

Introduction

1.1 Background

Sign languages are full visual languages, which make use of shapes of hands, hand motions, facial expressions and body language to convey a meaning [1]. They are used as the first language of millions of Deaf and hard-of-hearing people belonging to different parts of the world. But, not all sign languages are universal [2], [1] and every region has a particular system that is influenced by the cultural and linguistic factors. As an illustration, American Sign Language (ASL), Indian Sign Language (ISL), and Bangladeshi Sign Language (BdSL) have a great deal of differences in grammar, words, and composition [2]. The three languages were selected to conduct this study as they are used widely, and loanwords that are discussed in this work have alphabet-based datasets that are appropriate in the context of research on image classification.

Over 466 million people in the world currently experience disabling hearing loss and this figure is expected to rise shortly in the next decades [3], [4], [5]. In Bangladesh, one of the countries, hundreds of thousands of people use BdSL in their daily communication [6] . Although sign languages are quite common, Deaf people tend to have difficulties in communication in schools, hospitals, emergency services, and during an interaction with the rest of the population since not all trained interpreters are available [6]. Based on these challenges, there is a need to have dependable and affordable technological solutions that can be used to minimize communication gaps.

The recent developments in deep learning and computer vision made the automated sign language recognition promising. Convolutional neural networks (CNNs) have exhibited good performance in image tasks due to the ability to learn discriminative features of the visual image including hand shapes and minor variations of fingers [2] . ResNet and EfficientNet are the models that are now popular because of their predictable training behavior and the ability to learn meaningful patterns on images with high accuracy. Their effectiveness in classifying im-

ages render them the right candidates in noticing alphabets of the sign language that are in motion.

This study takes these advancements to discuss a multi-lingual model of recognition that can detect the signs of BdSL, ASL, and ISL. The research compares the results of ResNet and EfficientNet on three language datasets in these three languages and explores to what extent a single method can be applicable to other sign language systems. This background gives the impetus to the development of multilingual sign language translation aids based on deep learning and preconditions the further expansion of the arsenal in terms of providing communication assistance in real time.

1.2 Problem Statement

The individuals who are dependent on sign language usually have barriers to communication when communicating with people who do not comprehend their signs. Such obstacles influence most of the key activities such as education, health services, government services and emergency communication [6]. It is even more difficult in the case of multilingual settings where other sign languages are employed, e.g., BdSL in Bangladesh and ASL abroad or ISL in the local area [2]. Despite sign languages being a full-fledged visual language in the case of Deaf community [1] , the majority of the existing recognition systems are restricted to one language, which is rather restrictive in real-life contexts when there is a need to communicate in more than one language.

The system is in need of a lightweight and computationally efficient image based recognition system capable of detecting signs in more than one language with a common approach. Nevertheless, it is difficult to construct such system because of differences in gesture patterns, discrepancy in data quality, lighting effects and visual similarity of the shapes of hands in different languages [3] . To deal with such variations effectively, it is necessary to choose deep learning models that could extract meaningful visual features of different sign datasets.

The proposed research seeks to solve these problems by working on the creation of a multilingual sign language recognition system based on two popular image classification models, ResNet and EfficientNet. It is aimed at assessing the capacity of these models to identify indications of BdSL, ASL, and ISL and to learn their advantage and disadvantage in a bilingual environment [7].

1.3 Objectives

This project is primarily aimed at creating a multilingual sign language recognition system based on deep learning models on images. The following objectives are specific to the study to

attain the goal:

• To develop a single recognition system capable of categorizing hand signs of the Bangla Sign Language (BdSL), American Sign Language (ASL) and Indian Sign language (ISL).

• To build and arrange data on multilingual sign language, image processing (preprocessing and normalization) as well as adequate partitioning (training, validation, testing) are needed.

• To apply and test two deep learning models, ResNet, and EfficientNet, to test their effectiveness in multilingual sign language classification tasks.

• To compare the relative accuracy, stability and general performance of the two models under the same experimental circumstances to know their respective strengths.

• To determine the advantage and the drawback of image based classification toward multilingual sign language recognition.

1.4 Scopes

In this project, the author is planning to work on the creation of an image-based system capable of identifying and recognizing hand signs of the Bangla Sign Language (BdSL), American Sign Language (ASL), and Indian Sign Language (ISL) as a result of deep learning. The experiment is confined to the images of alphabet-level hand signs at rest, and all the classification tasks are conducted with the help of the ResNet and EfficientNet models. The datasets applied in this study were gathered, processed, and formatted to be trained, validated, and tested under this specific area.

There is no real-time detection, constant gesture detection and dynamic motion analysis or translation of complete sentences in the project. Neither does it deal with signer recognition, hand tracking or variance in representation by other viewpoints of the camera. Rather, the research will focus on assessing the effectiveness of the choice of models in recognizing multilingual hand signs in regulated image-based testing settings. The scope also involves training the models, comparing their accuracy, and their stability, preparing balanced datasets and making preprocessing steps.

Through these tasks, the project would grasp the advantages and drawbacks of image-based deep learning models, to multilingual sign language recognition. The results of the current work will assist in directing further studies on real-time systems, video-based recognition, and multilingual translation systems of large scale.

Chapter 2

Literature Review

2.1 Related Works

Sign language recognition has grown massively in the last decade with deep learning models being highly effective in arriving at visual patterns in hand gestures. Initial researchers were forced to use conventional machine learning and manually created features, but the advent of convolutional neural networks (CNNs) allowed achieving impressive gains in classification accuracy. Much of the available literature has concentrated on one sign language and researchers have come up with isolated solutions to BdSL, ASL, or ISL instead of multilingual systems. The subsequent review is a summary of the contributions that have been made and the gap that will inspire this research.

Image-based recognition has been studied in a few studies of Bangladesh Sign Language, and they employ CNN models. Hadiuzzaman et al [8] presented the BdSL47 dataset which owns the RGB and depth pictures of all BdSL alphabets and showed promising performance with the CNN-based classifiers [7]. Their work defined the significance of size of data set, image quality as well as diversity of signers to obtain credible results. Other studies on BdSL investigated how it can be applied in the real world, including real-time translation and sentence construction, and the growing necessity of precise and implementable BdSL recognition systems [6].These experiments demonstrate that CNNs are capable of learning pertinent features of hands in the case of a still image, however, it also indicates obstacles in the event of a gesture possessing a slight finger-level variation.

The study of American Sign Language (ASL) is relatively larger because of the accessibility of big groups of data. Variants of CNN are used in many ASL works to group alphabet signs based on image input or depth input. To illustrate that, more complex architectures like VGG or ResNet have been used in several studies to capture finer details of variations in hand shapes [2]. These models always do better than the shallow CNNs, as the deeper layers can learn hierarchical representations of features. Diversity of the dataset is also important in the studies of

ASL, where gestures differ greatly among signers, and across lighting and perspective. Despite the good results obtained in these works, they are limited to one language and do not cover the cross-language recognition.

This has been the case with Indian Sign Language (ISL) research. Most ISL systems make use of the imagebased datasets that are taken in the controlled laboratory settings, and the CNNs are trained to identify single gestures or alphabats. There are studies which have applied architectures allied to ResNet to detect ISL signs and have shown that more complex hand shapes can be better handled using deeper networks than shallow models [1]. Nevertheless, ISL datasets are generally smaller, and performance tends to reduce in case models are evaluated on images taken with varied light or camera configurations. These constraints make it clear that models that can be generalized to different conditions are required.

Deep learning has continued to provide great potential across all three languages: BdSL, ASL and ISL. ResNet and EfficientNet models have gained popularity in the models and are regarded as strong with regard to extracting features. The skip connections of ResNet also prevent the vanishing gradients in the training of deep networks, which allows the learning of detailed visual features to occur effectively [3]. EfficientNet on the other hand scales depth width and resolution equally across and thus can be trained to be competitive with much lower parameters [4] . Comparison of CNN versions has consistently demonstrated that more elaborate or better scaled architectures achieve high classification accuracy, especially on the more subtle gestures that require delicate finger movement, or on gestures that resemble each other in shape. In spite of these developments, the majority of the existing works only identify monolingually.

The systems that are trained on BdSL, ASL, or ISL only cannot be used in case of a multilingual environment, where the styles of gestures, convention of signing, and the structure of the alphabet differs significantly. Some recent studies tried cross-language training through combining dataset or using transfer learning methods but face various challenges such as dataset asymmetry, the appearance of gestures and inconsistent image quality [7]. Interference factors like variation in backgrounds, difference in hand size and variation by signers also complicate the process of developing uniform recognition systems. Multilingual sign language recognition is an area where there is still a gap in research. The current research on BdSL, ASL, and ISL proves that deep learning is effective in solving singlelanguage tasks, however, no scalable method is reported to solve unified recognition across languages. Based on the prior knowledge, this project compares two contemporary deep learning models, namely ResNet and EfficientNet, to three distinct sign languages in one system. The research has provided insights into the performance of these models on multilingual data, which can help in the creation of more general-purposive and inclusive sign language recognisers that can be used to accommodate the wider communication requirements.

Other research papers have examined the difficulties in developing effective sign language recognition systems that can serve different users. A common finding in BdSL, ASL, and ISL studies is that intra-class variation is very high, i.e. the same sign might look different when used by different signers, depending on the shape of their hands, their angle of attack, their skin colour or the surroundings. Some BdSL research observations were that models are not always capable

of differentiating between gestures with small rotations of fingers or those with similar shapes [6]. It was found that identical observations also apply to ASL works, where they find that deeper architectures like ResNet offer better learning by learning finer-scale spatial patterns that simpler CNNs cannot learn [2]. These results support the relevance of powerful feature extractors in the context of similar hand gestures that are visually related to each other.

The other major finding in the literature is that the quality and variety of the datasets affect the performance of the models. Most of the previous researches on BdSL and ISL used the small datasets that were captured in controlled settings and so the trained models had low generalization potential. Investigations emphasized that even the highly accurate models failed to perform well when they were tested on pictures of different backgrounds, where there was variation in lighting intensities or where the hand orientation was not used in the training material [1]. This drawback directly stimulates the necessity of multilingual data sets and different image samples. The ability of EfficientNet to scale up when changing and scaling its architecture, combined with its residual learning structure, has additionally granted ResNet-based classifiers a greater resilience to noisy or changing datasets [3], whereas the compound scaling of EfficientNet has maintained a consistent performance even when datasets vary significantly in terms of resolution or quality [4].

These are the strengths that make the two models appropriate in multilingual sign recognition. Applied to ASL and ISL, the impact of transfer learning was also investigated by researchers. Base models typically included the pretrained CNNs, like ResNet and EfficientNet, and retraining of classification layers on sign language images. This method saves a lot of time in training and is also more accurate, particularly when the size of the datasets is small. Various researchers found that transfer learning methods are better than models trained directly since they already contain pretrained weights that have general visual patterns which can be transferred to hand shapes and gesture features. Such findings indicate that the current architectures are capable of delivering high performance with relatively small datasets under the condition that the training process is organized and is backed with high-quality annotations.

The sign recognition research of cross-language sign recognition is minimal but provides resourceful insights to this study. Some recent studies have tried to combine datasets of various sign languages to determine whether the CNNs could make cross-linguistic predictions between languages whose alphabets and gesture shapes were different. These researches identified that although other models like the ResNet are capable of identifying certain cross-language gestures, the accuracy tends to decrease when the feature changes are substantial, or when the alphabets are based on contrasting structure rules [7]. This implies that multilingual recognition can indeed be achieved but with models that are able to acquire language specific and shared gesture features. The ability to scale effectively makes EfficientNet to be reported to be efficient in multilingual environment thanks to its balanced structure and reasonable use of the parameters [4]. These capacities closely go in line with the development objectives of having a single multilingual system.

There is also an increasing literature on the practical significance of constructing systems that facilitate the use of multiple sign languages, as opposed to the construction of individual models

of each language. The real-life user can be expected to come in contact with other users who belong to different language groups, particularly when attending schools and colleges, and also throughout the online communications. The users of BdSL in Bangladesh, in their turn, are often exposed to ASL signs in online content, and the users of ISL can communicate in BdSL or ASL in local or global community. Such situations cannot be covered by the existing single language models as they lack generalization outside their training language [6]. This drawback makes the argument of multilingual structures more compelling and encourages more studies on models that can be trained to use multiple alphabets with a single training paradigm.

Lastly, various articles talk about the need of the computational efficiency of deploying a sign recognition system to a real world setting. High-parameter models can be very good at accuracy, however can be unusable in low-resource environments such as a mobile phone, or other South Asian settings with limited resources. The literature has repeatedly pointed out the desirable tradeoff between accuracy and the cost of computation as evidence of the adequacy of EfficientNet [4]. Its architecture is scalable thus appropriate in light applications such as smart phone-based sign language recognition. Although ResNet is computationally more demanding, it is still stable in learning and reliable in feature extraction and can therefore be used as a formidable benchmark when analysing the performance of different models across languages. Through a comparison of the two architectures, past researches afford great information to the choice of models that are accurate and practical in application.

Altogether, the current literature proves that the single-language sign recognition is already on a high level, but also indicates the definite gaps in the field of multilingual studies. The vast majority of researches are also restricted to BdSL, ASL, or ISL alone and do not even seek to train or assess models that are capable of identifying gestures of more than one sign language at a time. There are very few works that apply ResNet or EfficientNet to do unified multilingual classification, and none of them united all three languages: BdSL, ASL, and ISL in a single framework. This is the gap that the current research is based on. Through the comparison of ResNet and EfficientNet on the basis of multilingual data, this study can make a new contribution to the model performance in cross-linguistic settings and promote the creation of more inclusive and integrated sign language recognition platforms.

Chapter 3

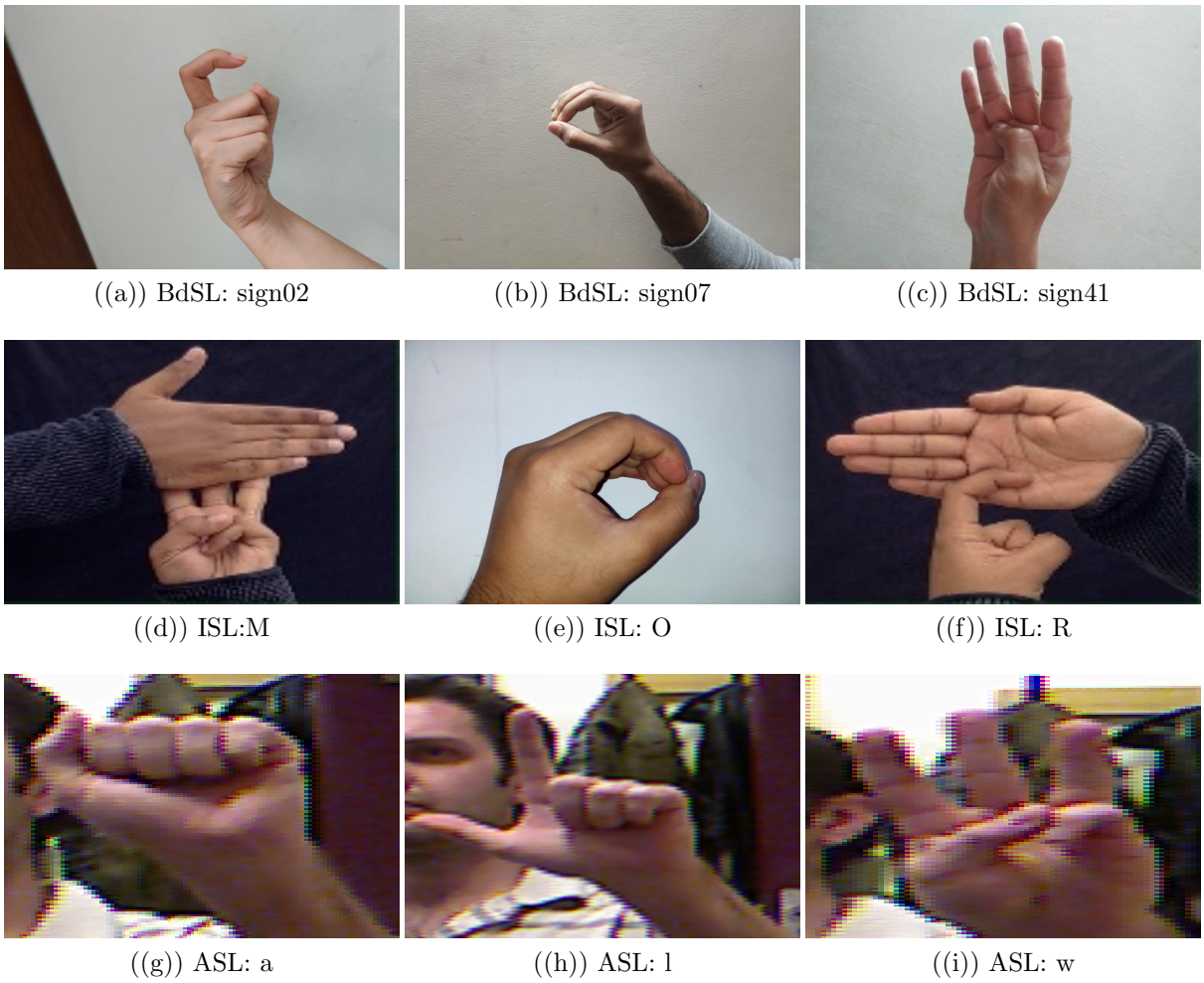
Dataset Description

3.1 Datasets

The datasets of this work were gathered in the most popular, open sources on Kaggle. These data sets are samples of hand signs in three languages of signs: the Bangla Sign Language (BdSL) [9], the American Sign Language (ASL) [10], and the Indian Sign Language (ISL) [11]. As the main goal of the current study is to create a multilingual sign language recognition system, it was important to choose the datasets that represent different geographical areas and methods of signing. The databases are unique in terms of the characteristics of their images such as image resolution, complexity of the background, orientation of the hands, posture, and the lighting circumstances. This diversity can assist the trained models to acquire strong and generalized features and hence limited reliance on a single dataset or controlled space and enhanced performance in the real-world. Each dataset has sample images as shown in Fig. 3.1.

3.1.1 Dataset Overview

- **BdSL:** The data of BdSL consists of the total number of 47,000 RGB images that depict the 47 sign classes that comprise 10 digits and 37 letters and were made by 10 different signers. The dataset represents a great range of the variations of hand shape, skin tones, hand size, signing angle, and orientation, as well as lighting and background environments. Although there were controlled and indoor-based lighting, there are also those that have natural and uneven light conditions. All pictures were reduced to 640 x 480 pixels.
  Due to the presence of many signers, the dataset has diversity in terms of hand appearance and signing style and would be very useful in training models that can work well in the actual conditions of the real world. This dataset can serve as a good starting point even though ASL is younger than BdSL datasets, as alphabet and digit-level classification can be performed with the help of this dataset.dataset provides a strong foundation for



((a)) BdSL: sign02 ((b)) BdSL: sign07 ((c)) BdSL: sign41

((d)) ISL:M ((e)) ISL: O ((f)) ISL: R

((g)) ASL: a ((h)) ASL: l ((i)) ASL: w

Figure 3.1: Sample images from the three sign language datasets used in this study. Each row shows three signs with their corresponding alphabet meaning: BdSL (top row), ISL (second row), and ASL (bottom row).

alphabet- and digit-level classification.

- **ASL:** The ASL Fingerspelling dataset is composed of RGB and depth images of 24 letters of the English alphabet (except X or Z). Five non-native signers were used to collect the data using a Microsoft Kinet sensor. Despite the availability of depth information, this study only used the RGB images in this study to ensure that there was consistency in all the three languages used.
  ASL datasets tend to be larger and more standardized since they have been widely used in the world research. The photographs in this project feature the samples of various people and are filmed slightly different viewing angles. These variations enable the model to be learned to understand that the same sign can be represented differently by varying hand orientation, finger placement and camera angle which enhances recognition strength.

- **ISL:** There are about 1,200 images per class in the ISL dataset of 35 handshapes of the Indian Sign Language. The data were collected over several external and self-created sources leading to high variance in the lighting conditions, the complexity of the back-

ground, and the shapes of hands, and signer styles.
ISL images have heavy shadows and complicated backgrounds, in comparison with BdSL and ASL, and create further training challenges. Nonetheless, it also adds value to the dataset in terms of building a model that is resistant to environmental change. The fact that the use of ISL has a large populace in India means that the dataset has a natural pull to diverse range of signers and hand shapes, which will increase the ability of the model to extract features.

3.1.2 Dataset Collection

The data utilized in this study was gathered through publicly available sources and three big sign languages, including Bangla Sign Language (BdSL), American Sign Language (ASL), and Indian Sign Language (ISL). The individual datasets were initially generated in varying acquisition conditions and experimental configurations which added natural diversity in terms of signer traits, background conditions as well as quality of the images. This heterogeneity was necessary to create a strong multilingual system of sign language recognition:

- **BdSL:** The BdSL data comprises a very high amount of images that encompass the Bangla alphabet and numbers. The photographs were obtained among various signers, thus, the images contained natural diversity of skin tone, hand form, size of hands, angles of signing, and speed of motions. The controlled indoor light conditions were used to capture some of the samples whereas the nature and uneven light conditions were used to capture the rest. Such variation makes the dataset more complicated and more useful in that the model was trained to be able to provide responses to irregularities in the real world. BdSL datasets are quite new in comparison with the ASL ones, but they offer a robust basis on the classification of alphabets and figures.
- **ASL:** The ASL dataset consists of structured pictures of the English alphabet which are gathered with the help of a Microsoft Kinet device which records both color and depth data. The signs were repeated more than once by five non-native signers, which made the dataset record the minor changes in the hand orientation, finger positioning, and the posture of signing. The RGB images were utilized in this research although the data of depth were also available to facilitate the case of consistency in all three sign languages. The ASL datasets are usually larger and more uniform because of their huge worldwide research application and the standardized conditions of acquisition.
- **ISL:** The ISL dataset has been made up of images depicting the ISL alphabet and common used hand gestures. In contrast to BdSL and ASL, the ISL dataset also covers more complicated backgrounds, more prominent shadows and variable conditions of light, which is more difficult yet more productive in enhancing the robustness of the models. The data set shows a broad combination of hand shapes and finger positions as the signers

are diverse from a large population in India. This has a broad range of variation, which adds more feature extraction and generalization to the model.

3.1.3 Dataset Preparation

After collecting the datasets, the second task was to prepare the datasets in a way in which the deep learning models could learn successfully. As the pictures were the product of various sources of the masses, and were not organized in a single structure, it was necessary to prepare the data sets in order to achieve uniformity and trustworthiness. This phase entailed cleaning of images, resizing, organization of folders and labeling of classes.
The initial stage of preparing the data was the standardization of the size of images. The images in the original datasets had different resolutions, thus, all of them were re-sized to a common resolution of 224 x 224 pixels to fit the input specifications of the ResNet and EfficientNet models. By having a consistent image size, one is able to extract features and also makes it compatible to all the experiments. Also, all images were converted to the RGB color format to remove any discrepancies due to the differences in grayscale or multi-channel differences.

This was followed by pixel value normalization. The pixel intensities were divided by 255 to bring them to the range [0, 1]. This action increases the stability of the training, convergence, and overall generalization. Given that the datasets were sampled in varying lighting conditions, normalization was also useful in eliminating illumination variations.

After some basic preprocessing was done, much work was devoted to the annotation of the datasets and reorganization of folders. Due to the fact that the different contributors have developed each dataset, there was inconsistency in the naming conventions, folders hierarchy and class labeling formats. All folders were renamed in regards to the real alphabet or gesture names to make the training pipeline clean. This was a highly crucial step since the model uses the name of folders as a direct determinant of class identities.

To further enhance the quality of the dataset, all extraneous files such as text files, damaged images and misclassified samples were eliminated. Manual inspection was also done carefully to determine and eliminate the wrong or noisy samples. This cleaning procedure greatly decreased label noise and raised dataset reliability.

3.1.4 Data Preprocessing

Following the preparation and annotation of the datasets, more advanced preprocessing was done to enhance model performance. Through a common preprocessing pipeline, all images were processed to guarantee that the multilingual dataset was homogeneous.

The preprocessing steps included:

Image Resizing: All images were resized to 224 × 224 pixels.

Normalization: Pixel values were scaled to [0, 1].

Color Standardization: All images were maintained in RGB format.

These measures were to make sure that the deep learning models are presented with a standardized input irrespective of the origin of the data. This standardization is necessary to the objective training and it is more so in a multilingual classification setting.

3.1.5 Data Balancing and Preparation

Once preprocessing and annotation were completed, the other significant step was balancing the dataset, and preparing it to be used in training the model. The samples per class was very uneven considering the fact that the images were obtained in various sources. There were alphabets that had hundreds of images and some with few dozen images. Such unbalanced data would be subject to training and influence the model towards higher sample classes and adversely impact fairness and general performance.

In order to solve this problem, a balancing approach to classes was used. A certain amount of images was picked within each class such that there was an equal number of training samples in each class. This process decreased the number of images, but formed an equal, smooth, and objective dataset of all the three sign languages.

Once the dataset was balanced, the images were divided into three subsets:

- Training set: Used to train the model and learn feature patterns.
- Validation set: Used to monitor training performance and detect overfitting.
- Testing set: Used only for final performance evaluation on unseen data.

Images were randomly mixed and split to avoid the ordering effect. This was done so that the samples of BdSL, ASL and ISL were distributed equally in all subsets..

The dataset was organized to a clean directory hierarchy to support the efficient training and evaluation. All the sign languages and words were maintained in distinctly marked folders within their respective training, validation and test folders.

Conclusively, data balancing and preparation stage was vital in providing fairness, stability and accuracy of the multilingual sign language classification system. These procedures were necessary to guarantee that the input data provided to ResNet and EfficientNet models were of high quality, balanced, and well-structured, thus allowing to conduct reliable performance assessment and make sense of model comparison.

Chapter 4

Methodology

4.1 List of tools used

This project involved the use of both hardware and software to prepare the data, train the models and analyze the results. The tools were chosen according to availability, ease to use, and their capabilities to be used in deep learning and image processing applications.

4.1.1 Hardware Specifications

All the models are trained and tested in Google Colab [12]. The hardware specifications of the experimental setup used in this study are summarized in Table 4.1.

Table 4.1: List of Hardware Specifications

Parameter	**Specification (Google Colab)**
CPU Model	Intel(R) Xeon(R)
CPU Frequency	2.20–2.30 GHz
Number of CPU Cores	2
GPU Model	NVIDIA Tesla T4 / P100
GPU Memory	12–16 GB
RAM	12 GB

4.1.2 Software Specifications

The software libraries and tools used for model development, training, and evaluation are summarized in Table 4.2.

Table 4.2: Software Libraries and Tools Used in This Study

Library / Tool	Generic Name	Version
NumPy	numpy	1.23+
Pandas	pandas	1.5+
TensorFlow	tensorflow	2.12.0
Keras	keras	2.12+
PyTorch	torch	2.1
OpenCV	cv2	4.7.0
Matplotlib	matplotlib	3.7+
Scikit-learn	sklearn	1.2+

The experiments were conducted using NumPy [13], Pandas [14], TensorFlow [15], Keras [16], PyTorch [17], OpenCV [18], Matplotlib [19], and Scikit-learn [20].

4.2 Methods

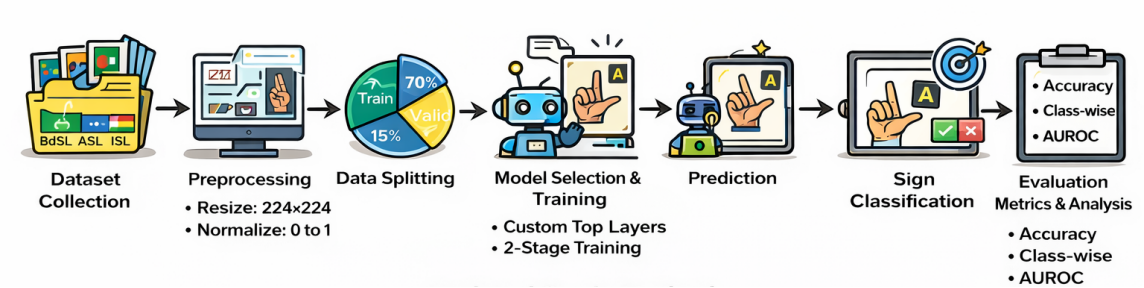


Figure 4.1: Research workflow showing the main steps followed in this study.

The entire sequence of the present research can be presented in a step-by-step diagram, which is displayed in Fig. 4.1. It begins with the gathering of datasets of four languages of signs, then preprocessing and model training. The subsequent steps are prediction, translation to target language and evaluation of performance. This diagram gives a brief description of how this study will be carried out. Each step of the proposed methodology will be described briefly in the following sections.

4.2.1 Data Loading and Sorting

Here, the datasets were ready in such a manner that it could be utilized by the deep learning models. Because, in our study, we had to deal with several sign languages, namely, the Bengali (BdSL) one, the American (ASL) one, and the Indian (ISL) one, the number of classes in each of the folders was different. The loading and sorting process had to be consistent in order to prevent a mis-matching of images and labels. To start with, all image paths were loaded using their data set directories. The folder names were used as the labels of our classes since each

one was named after a given sign. Once the paths of the images were acquired, we tried to arrange the paths in alphabetical order so that the preprocessing process would maintain the same order between the path and the object categories. Once the images were loaded, files were changed to RGB format (instead of the default OpenCV format of BGR) in order to have the same color representation in all files. The images were resized to 224x224 pixels, the input dimension needed by both ResNet and EfficientNet models, and then all the images were cut down to the same size. This sorted load and sorting made sure that: Every image was paired with the correct class label. Entries were not duplicated or lost. Each set of data had the same format. During training, models used clean and consistent input This was done to ensure that data was consistent and errors were minimized when developing a model.

4.2.2 Normalization

Once the images were loaded, the next thing that was done was to normalize the pixel values, and thus the datasets all fell within the same numerical range. The models can also learn slowly or randomly without prior scaling of the raw images because pixel values, which are present in the raw images, range between 0 and 255. To overcome this, all the images were scaled by dividing the pixel values with 255. This normalized the whole data to the 0-1 range which is appropriate when using deep learning networks like ResNet and EfficientNet. The use of normalization allowed the models to train more effortlessly, minimized the chances of numerical instability, and guaranteed that all the datasets had consistent values of the input during training, validation, and testing.

4.2.3 Splitting the Dataset

We divide our dataset into two subsets a training set and a test set that we used to train and evaluate our model 80 percent of the data is in the training set it has both the impressions images and the mask images of the images. This subset trains the model, which in turn helps it to learn by the patterns and correlations of the data. We separated the data with a train-test split method of the scikit-learn library [21] with the random state parameter value of 42. The random state that we use is to ensure that our split dataset is consistent across multiple runs of the code. This means that the training and test set will be identical in the event of a repeated execution of the code with the same random state value (42), the results will be identical. 20 percent of the entire data set was used as the test set. This set contains impression and mask pictures that have not been put on during training. The test set is used to test the effectiveness of our trained model on new data..

4.2.4 Deep Learning Models (ResNet & EfficientNet)

Multilingual sign language images need a model that can comprehend the advanced visual patterns, hand shapes variations, background variations, lighting variations, and variations within a given class in Bangla, American, and Indian sign languages. To do so, two convolutional neural network architectures ResNet50 and EfficientNetB0, which are state-of-the-art, were chosen to be used in the experiment. Both models are known to be very powerful in terms of feature extraction and are complements of each other, since they are based on two design philosophies, which are depth-oriented with residual connections and balanced-scaling based with computational efficiency. ResNet50 (Residual Network, 50 layers version) is another type of deep architecture which brought the idea of residual learning. With the classical deep networks, the gradient decreases in size with the increased number of layers, and hence the model ceases to learn. ResNet helps to address this problem, introducing identity skip connections that enable the gradient to pass through several layers without information loss. This helps the network to acquire more complex patterns with constant training.

The starting ResNet50 model of this study is already trained on ImageNet that consists of millions of labeled images. These pre-trained weights facilitated a fast learning process by the model because it does not need to learn low-level features at all. The last fully connected block of the ResNet50 was also removed and the classifier replaced by a custom classification head corresponding to the amount of classes of signs in the combined datasets. The custom layers were a global average pooling layer to down-sample the feature maps, a dense layer to interpret the features, batch normalization to stabilize training, dropout to regularize the model, and final softmax to generate the probabilities of the classes. The second model applied is EfficientNetB0, which is more a contemporary approach towards network design. EfficientNet does not scale the network merely by adding depth or width but in a balanced way with the help of a compound coefficient. This scaling technique enables the model to be efficient and yet attain competition. EfficientNetB0 is not that heavyweight as ResNet50, but it is reported to be better at fine-grained features with a lesser amount of parameters. Like ResNet, EfficientNetB0 was pre-trained on ImageNet weights and changed by overriding the top layer with a multilingual sign language recognition classification block. The model was designed to cope with the differences in the hand orientation, subtle finger position and the visual style of three country sign languages. The two models had a two stage training procedure. In the first phase all the pre-trained layers were frozen in order to update only the newly added classification layers. This enabled the model to adapt to the new world of sign language images and not to disrupt the acquired visual features. When the new layers had a stable performance, a few layers of the base models were unfrozen and fine-tuned. Fine-tuning gave the models the capability to pick up additional patterns that were more specific in terms of hand shapes and gestures that are particular to BdSL, ASL, and ISL. Adam optimizer was used in training at a carefully selected learning rate that ensured that the models did not exceed the minima. Stopping Training was Early termination of training was done when no more improvement was being made and a learning rate scheduler was applied when validation performance started to stagnate.

During the training process, both EfficientNetB0 and ResNet50 were trained on the same dataset splits so as to be equal in the comparison. ResNet50 was deeper and tended to capture more abstract features whereas EfficientNetB0 was better in terms of computational efficiency and faster convergence. Analyzing these two models, the given research has come to understand which architecture is better to apply in case of multilingual sign language classification and how various design principles influence the performance in case of working with different sets of gestures.

4.2.5 Sign Translation Pipeline

The sign translation pipeline consists of three stages:

1. **Sign Prediction:** Input images are passed through the trained classifier to identify the corresponding sign. For continuous gestures, predictions across frames are combined to form a complete word.
2. **Text Translation:** The predicted word is translated into the target language using the Google Translate API, enabling multilingual communication.
3. **Target Sign Mapping:** The translated text is mapped to corresponding sign images in the target sign language dataset (e.g., BdSL, ISL, ASL). This step provides visual output for users who rely on sign language.

4.2.6 Evaluation Metrics

Model performance was evaluated using classification metrics derived from the confusion matrix, including Accuracy, Precision, Recall, F1-Score, and AUROC.

Accuracy measures the proportion of correctly predicted samples:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

Precision measures the ability to avoid false positives:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.2}$$

Recall quantifies the ability to detect true positives:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.3}$$

F1-Score, the harmonic mean of precision and recall:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.4}$$

AUROC measures the separability of classes across decision thresholds.

The **confusion matrix** provides a detailed view of class wise predictions, revealing which signs tend to be misclassified. These evaluation metrics allow comprehensive comparison of two models.

4.2.7 Generalization Strategy

It is important that a sign language recognition model is reliably good on unseen data as to all users there are natural variations between the different users of the sign language, hand shapes, and recording environments. There were no specific data augmentation methods used in this research, but there were a few of the effective strategies that were employed to improve the generalization ability of the model.

To begin with, a balanced dataset splitting strategy was used, in order to ensure equal distribution of classes among the training, validation, and testing sets. This avoided biasness to any of the classes of gestures and made the performance assessment represent the actual capacity of the model to generalize to previously unknown samples.

Secondly, the model structure and training parameters were chosen very carefully to reduce overfitting. Regularization measures like dropout layers and early stopping were also added to stabilize the learning process and limit the model to memorization of particular patterns in the training data. These mechanisms stimulated the network to concentrate more on the more general and discriminative feature representations.

Lastly, a totally independent test set, obtained using datasets that had never been used in training or validation, was used in the evaluation of the model. Such rigid division provided a realistic and objective evaluation of the model performance in the real world.

Chapter 5

Results and Analysis

Having trained both ResNet50 and EfficientNetB0, it was necessary to study the results in detail to determine the performance of the models in working with the multilingual sign language samples. This analysis aimed at not only measuring the accuracy, but also to note the behaviour of each model training and testing, whether the models were able to learn the patterns in the data in a consistent fashion, and where the differences between the two architectures appeared. The evaluation process allowed determining which model was more responsive to the changes in the data as machine learning models tend to exhibit varying degrees of sensitivity to the changes in data. The two models performed some promising results during the experiments although they displayed different learning behaviors. ResNet50 used as the deeper residual architecture showed consistent gains throughout the epochs. The training curve was smooth and slow to learn and the validation curve was near to the training performance, which means that the model was overfitting it badly. EfficientNetB0 was a smaller model and thus was convergent in a shorter period. Its training performance improved faster at the initial stages, but validation performance was sometimes slightly fluctuated. Despite such changes, EfficientNetB0 still performed highly, and the training took shorter hours, which is indicative of its efficiency-driven architecture. Table 5.1 shows the performance of EfficientNet-B0 and ResNet-50 in terms of their macro-average performance metrics on the test dataset.

Table 5.1: Macro-average metrics for both models on the test dataset

Model	**Precision**	**Recall**	**F1-score**	**Accuracy**
EfficientNet-B0	0.94	0.94	0.94	0.94
ResNet-50	0.96	0.96	0.96	0.96

The final test results were assessed, and it was revealed that ResNet50 is slightly more effective in complex classes where the gestures were characterized by slight visual differences. Efficient-NetB0, in its turn, has given a very competitive result with a reduced number of required computational resources. The variations in the outcomes were used to bring out the strengths of either architecture. ResNet50 proved to be more useful in getting finer grained features of

the images, whereas EfficientNetB0 was faster and lighter and it could be used in real-time applications. On the whole, the discussion of the finding of the experiment indicates that both models were able to learn to classify signs of the Bangla, American, and Indian sign languages. The comparison made between the two performance patterns was very helpful in giving the insight of the effect of architecture on the model behavior which was subsequently used to give a more detailed judgment during the subsequent subsections.

5.1 Model Performance Evaluation

The models were assessed in terms of their ability to classify the images (both in the training and validation stages) correctly. The dynamics of the accuracy and loss curves were also significant in all the experiments to establish the extent to which the models were adjusting successfully to the multilingual sign language datasets. In the case of ResNet50, the accuracy of training improved step by step, and the learning process was considered stable. Similar pattern was observed in the validation accuracy, which remained near the training accuracy, which means that the model was not overfitting and learning meaning features. The loss values of ResNet50 decreased with time, which again supports the fact that the model was learning to extract useful representations of the images and as the model continued to train, the loss values were reducing.

The models were assessed in terms of their ability to classify the images (both in the training and validation stages) correctly. The dynamics of the accuracy and loss curves were also significant in all the experiments to establish the extent to which the models were adjusting successfully to the multilingual sign language datasets. In the case of ResNet50, the accuracy of training improved step by step, and the learning process was considered stable. Similar pattern was observed in the validation accuracy, which remained near the training accuracy, which means that the model was not overfitting and learning meaning features. The loss values of ResNet50 decreased with time, which again supports the fact that the model was learning to extract useful representations of the images and as the model continued to train, the loss values were reducing.

After the evaluation of the test results, both models demonstrated good classification abilities. ResNet50 had greater accuracy in the classes that involved the similarity of the gestures in terms of appearance and demanded additional feature extractions. EfficientNetB0 also performed remarkably well and it has only a smaller number of parameters but its level of accuracy is similar to ResNet50. The obtained results of the evaluation confirmed the fact that both models managed to learn the differences that exist within BdSL, ASL, and ISL, and that able them to recognize the forms of the complex hand and gesture patterns. The similarity in the performance levels during the training, validation, and testing phases indicated that the models were trained well and could effectively deal with multilingual sign language classification.

5.2 Analysis of Prediction Accuracy

The accuracy of prediction of the models was analyzed in order to know how well they identified the gestures in the three sign languages that were included in the dataset. Through the assessment of the correctness in training and testing, it was feasible to note the extent of each model to the variations in the hand shapes, orientations, and visual differences among BdSL, ASL, and ISL. ResNet50 was found to be very successful in this aspect especially due to its deep structure that enabled it to learn more detailed and abstract characteristics of the images. The results of the accuracy tests demonstrated that ResNet50 could detect small patterns like a small finger movement or even slight changes of hand position that are usually hard to see in multilingual sign language data. This additional level of feature extraction helped the model to achieve a high degree of accuracy in most of the classes of gestures.

EfficientNetB0 also had similar and steady accuracy but the behavior was a bit different as compared to ResNet50. The model was very effective when the gestures were clear and distinct visually during classes. Its balanced architecture contributed to its generalization to various lighting conditions and backgrounds, which were also found in some part of the dataset. But in classes where the signs bore very similarity or had a small variation among the signs, EfficientNetB0 performed slightly worse in terms of misclassification. Such minor drops in accuracy were to be expected because of the more compact shape of the model, which restricts the richness of its feature extraction relative to more profound networks such as ResNet50. Figure 5.1 shows the sample predictions of EfficientNet-B0 and ResNet-50 with their score of confidence and the ground-truth labels.



Figure 5.1: Predictions from EfficientNet-B0 and ResNet-50 with confidence scores and ground-truth labels.

The two models despite these differences performed well in terms of predictive accuracy and were also good in terms of dealing with the multilingual character of the dataset. The test results values of accuracy indicated that both models were able to identify gestures in all three sign languages and hence were able to cross cultural and visual differences in sign representation. The relative evaluation of their performance on prediction accuracy demonstrates the impact of architectural differences on the performance of the models, with ResNet50 being superior in those necessitating more profound pattern recognition and EfficientNetB0 being more efficient and faster to generalize.

5.3 Comparison Between ResNet and EfficientNet

ResNet50 was compared with EfficientNetB0 to gain an insight into the performance of the two architectures under identical settings and using identical multilingual sign language data. Despite the fact that both models yielded good results, their behavior during the training and testing process showed that they had observable differences that were affected by their structural design. ResNet50, which is a more profound residual network performed better in a scenario where fine-grained feature was required. It was also able to learn complex patterns, which enabled it to better deal with gestures that have minor variations. This was also captured in that it was more accurate in a number of closely related gesture classes meaning that the deeper layers made the model to capture fine visual details.

On the other hand, EfficientNetB0 was highly efficient during the training process. It needed less computation time and memory than ResNet50, and still could achieve competitive performance. The compound scaling strategy of the model enabled it to trade depth, width and resolution such that learning stability and quick convergence was obtained. In situations where gestures were very similar in shape or where greater hierarchical features were needed, EfficientNetB0, in some instances, demonstrated higher misclassifications than ResNet50. Such differences did not diminish its overall performance much, but they brought out the trade-off between the computation efficiency and the depth of representation. In figure 5.2, the accuracy of the classes with a threshold of less than 0.8 in the models of BdSL and ISL are represented. Figure 5.3 shows the mean macro AUROC of each of the four sign language datasets of all the models. The two figures 5.2 and 5.3 jointly demonstrate the class wise weaknesses and the discriminative performance in general of the proposed models.

After the comparison of the final test results, it was evident that the two models can address the issue of multilingual sign language classification successfully. ResNet50 managed a minor lead in total accuracy and stability and in more complicated gesture categories especially. EfficientNetB0 was almost equally effective, but it provided the advantage of being trained faster and using less resources, so it can be used in any application where efficiency is required. The analysis indicates that the decision between the two models would be determined by the needs

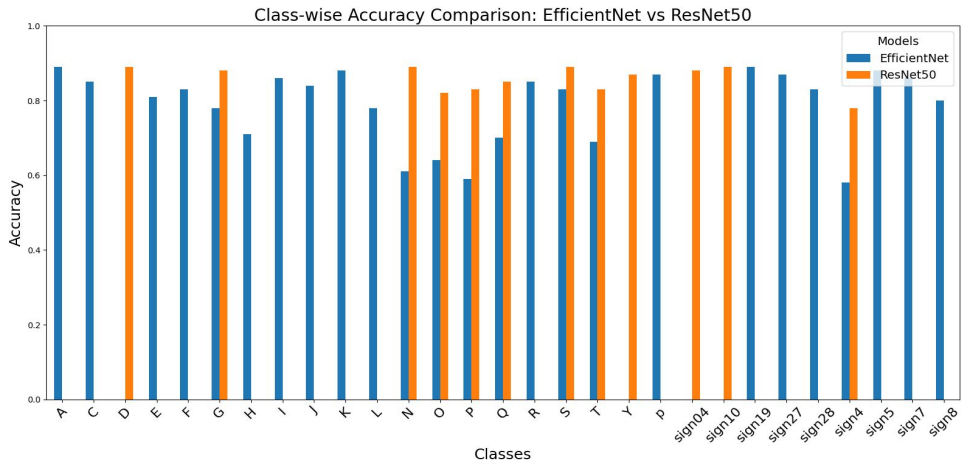


Figure 5.2: Classes with accuracy < 0.8 in BdSL and ISL for all models

of the targeted system. ResNet50 is stronger in case maximum accuracy is a requirement and there are enough resources in the form of hardware.

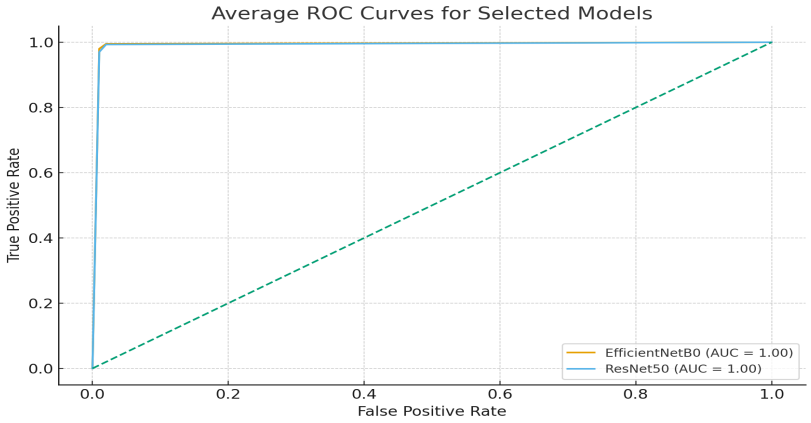


Figure 5.3: Average macro AUROC for all models across BdSL, ASL and ISL

Chapter 6

Social Impact

The importance of the classification system of a multilingual sign language is social, particularly in those countries when the barriers of communication still screen out the deaf and hard-of-hearing population of the community in educational, employment, and everyday life. Sign language users find it difficult to interact with other people, who do not understand their signing system, and it poses difficulties in receiving such necessary services as healthcare, banking, or transportation. With the introduction of a system that can identify gestures of Bangla, American, and Indian sign language, the present study will decrease these gaps and provide a technological transition between signers and non-signers. The fact that the system can be used in a variety of languages makes it more useful in various communities and improves the accessibility to various regions. The incorporation of deep learning architecture in social communication devices indicates the possibility of the contemporary technology to facilitate the inclusive development. The effective work of ResNet50 and EfficientNetB0 proves that sign language translation systems are constructable with high confidence and are affordable to everyone with not so much technical aspects. With the ever-growing digital service, it is possible to enhance communication within education, government, transport, and workplaces by having intelligent applications of gesture recognition. This technology has also influenced the creation of awareness of the requirements of the deaf community and the significance of learning the sign language as an inclusive communication method.

6.1 Sustainability of the Work

The sustainability of this research is that it can be effective in the long run, as it can be adjusted to meet the requirements of the future. As the system is supported with machine learning models, it can keep being enhanced once bigger or more varied datasets are accessible. The models to be used in this project, specifically EfficientNetB0, are lightweight and can be used in devices with low processing power and hence the system can be used in a variety

of environments in the long run. This helps in sustainability as it does not require high-end hardware to implement and also because the solution can be implemented in low-resource hardware like rural institutions, small schools or community centers. Moreover, the system is multilingual, which means that it will be able to cater to a broad audience and be useful at the time when the communication requirements will develop. The need to have convenient and reliable translation systems will also increase as the number of individuals who understand sign language and use digital devices in communication increases. The nature of the project ensures that the new sign languages, regional variations and new classes of gestures are easily added later in the future without the need to redesign the entire project since it is created on scalable deep learning techniques. This flexibility enhances the sustainability of the work in the long term and makes it possible that the system will be able to change with the technological and social changes.

6.2 Social and Environmental Impact Analysis

The technical process of creating multilingual recognition technologies of sign language has serious social and environmental connotations, especially in the area of assistive technology and accessibility.

6.2.1 Social Impact

Socially, our system of sign language translation focuses on dealing with contentious communication obstacles experienced by the deaf and the hard-of-hearing members. The technology facilitates social inclusion and equal participation since it allows the easy communication of all users of Sign language and non-signers. The fact that the system allows cross-lingual translation of the sign language also increases the intercultural communication of the global deaf community. Moreover, the technology will make sign language more accessible to non-deaf people, which will prompt more members of society to consider the deaf culture and language, thus possibly mitigating social isolation and initiating an understanding process.

6.2.2 Environmental Impact

The socially relevant implications of the introduction of a multilingual sign language recognition system are significant in that the situation will impact individuals, who rely on the sign language as their main form of communication. The system will promote interactions among signers and non-signers by offering a computer-built solution to interpret signs in Bangla, American, and Indian sign languages. This can bring about some good developments in schools, work place and service communities where misunderstandings will mostly arise due to the lack of communica-

tion. With the increasing use of technology-based communication devices by more organizations, hearing-impaired individuals can have improved access to key services and opportunities, which were otherwise hard to access. Through this, the system can facilitate equality and enhance social engagement of the deaf and the hard-of-hearing communities. Environmentally, the system does not use any physical materials or large-scale production to make it as the ecological footprint is reduced. Seeing that the models can be installed on the existing digital devices, the manufacturing of the additional hardware is not needed, which decreases the environmental footprint of the electronic waste. Training and deployment via the cloud also enable the system to execute via shared computing resources, which restrain the use of individual high-powered machines. Though, the training of deep learning models is energy-consuming process, it is conducted once and can be reused multiple times without incurring substantial energy demand. All in all, the work has a low environmental footprint, and the system provides a sustainable digital solution, which is in line with the current attempts to minimize the consumption of resources.

Social Effects

This study has a high social impact as it discusses the problem of communication of deaf and hard-of-hearing people. Through offering the system which is able to identify the signs of various sign languages, the work contributes to the comprehension of how to improve communication between signers and non-signers in real life situations. This will help them to be less socially isolated and foster more productive relations in schools, work places, hospitals and service centers. In case of easier communication, individuals with hearing disabilities are better placed to engage in academic and work related activities, and that makes people live in a more inclusive society. Through the system, people are also made aware of the need to use sign language and encouraged to learn and understand various signing systems so that more social acceptance and support can be achieved.

Environmental Effects

The environmental footprint of this work is not very high, since the system is based on fully digital processes instead of using materials or manufacturing. The trained models can also be run on the current machines like the computers or hand-held phones without hardware requirements. This lowers the manufacturing of electronic parts and contributes to the restriction of e-waste production. Dynamic deep learning models have a cost to run during the training stage, but during this stage, only once is required, and the trained model could be reused again and again with a minimal amount of extra energy cost. Cloud systems also maximize the utilization of resources by allowance of a single hard-copy to be used by many different users, minimize the impact on the environment that can be linked to having a dedicated computing system. In general, the project contributes to the eco-friendly strategy, using digital infrastructure and reducing the necessity to use physical resources.

6.3 Ethical Considerations and Framework

Ethical factors are a significant part of any given system that deals with the real users especially in communication and accessibility fields. Given that this is a work that sign language must be recognized, it is important to make sure that the system does not infringe on the privacy and dignity of the people whose gestures are under deciphering. The information in the datasets of the given research was publicly accessible and had no identifiable personal data that would promote the preservation of data privacy and minimized the ethical issues associated with consent.

The system is to help with communication rather than to substitute human interpreters and to disrupt personal communication, and this intention was the main consideration of the development process. Another fact to consider is being fair and non-biased when making predictions by the model. During training, the dataset included very different samples of various sign languages and this factor helped to minimize potential prejudice against a certain group. Nonetheless, the evaluation should be done on a regular basis to make sure that the system works in the same manner with individuals with varied skin tones, backgrounds and hand shapes. Implementation should be done ethically by bringing out the limitations of the system and using it responsibly in practical situations. Considerable accessibility can be improved with the help of the technology, without ethical values being undermined, promoting a more inclusive and respectful attitude to communicating with the deaf and hard-of-hearing population.

Chapter 7

Lesson Learned

Even during the progress of developing the multilingual sign language classification system, some essential lessons have been obtained that helped get a better idea of both technical and practical issues of creating an accessible tool of communication. This interaction with datasets of three sign languages demonstrated that the hand shapes, backgrounds, and lighting conditions can thus influence the work of the deep learning models. These differences had to be carefully preprocessed, balanced correctly in data and constantly evaluated in order to make sure that the correct patterns were learned by the models. This process revealed the significance of regular dataset preparation, as minor differences in labelling or images quality may cause wrong predictions and decrease accuracy.

The other lesson that I learnt during the project is the importance of selecting the correct model architecture to accomplish certain tasks. Despite the similar positive results of both ResNet50 and EfficientNetB0, their behavioral characteristics throughout the training process indicated the impact of architectural peculiarities on the learning rate, the generalization capacity, and computational demands. This knowledge will come into play in the future study, particularly when implementing similar systems in settings with limited resources. Besides, the project also stressed the need to check the performance of validation and use various approaches, including fine-tuning, early stopping, and learning rate scheduling, to ensure stability throughout training.

Lastly, the experience of working on this project also allowed seeing some bigger issues regarding the accessibility and communication of the deaf community. It is now apparent that technological solutions have to be presented in a responsible way, with ethical implications, user privacy and usability of the solution considered in real life. All of these lessons led to a more in-depth idea of how complicated the process of creating a multilingual sign language recognition system is and how essential planning, unceasing experimentation, and careful implementation may be.

7.1 Challenges Faced

In the design of this multilingual system of classifications of sign language, there were various issues which were faced that affected the flow of work and necessitated keen improvisations which would enable credible findings. The disparity in the three sign language datasets was one of the major issues to deal with. The datasets possessed their own peculiarities such as varied lighting conditions, background noise, resolutions of the image and styles of gestures. Such inconsistencies could at times lead to misinterpretation of some signs by the models and more preprocessing measures were required to reduce the effect of these differences. Another problem concerned the balanced representation of classes since there are gestures with much less samples than in other classes, which influenced the capacity of the model to learn these gestures.

The other important issue was the ability to handle the computational demands of the process of training deep learning models. As a deep architecture, ResNet50 was more time and memory consuming to train. The model required patience and constant observation despite the help of GPU-based settings to be fine-tuned to avoid excessive overfitting or erratic accuracy decline. EfficientNetB0, which was lighter, was sensitive to minor variations in hyperparameters, and multiple experiments needed to be re-run in order to reach consistent performance. It took a series of trials and critical assessment to determine the appropriate learning rate, batch size and training epochs. Lastly, the conceptual issues were also associated with the incorporation of multilingual datasets into a single system. The models were forced to acquire not only the peculiarities of each sign language but also the common trends between the sign languages. This required a balanced approach of training in which one dataset was not predominant in the learning. The struggles to cope with these challenges were good experience and contributed to the creation of a stronger and more flexible system to classify sign language.

The results obtained in this paper indicate that deep learning models are capable of identifying hand signs of the Bangla, American, and Indian sign languages with a deep learning input based on only the image data. ResNet exhibited high feature extraction especially in cases where the signs involved are similar in appearance whereas EfficientNet offered competitive accuracy at a reduced computational cost. These findings indicate some potential to use both models in multilingual sign language recognition systems as effective parts. Nevertheless, the general results are also about the significance of the diversity of datasets, regulated preprocesses, and balanced training samples when addressing multiple sign languages. The results of the study provide an excellent guidance on the future research, especially in the extension of the system to real time applications and multilingual larger data sets.

The findings of this study show that deep learning models can successfully recognize hand signs from Bangla, American, and Indian sign languages using only image-based inputs. ResNet demonstrated strong feature extraction capabilities, especially when dealing with visually similar signs, while EfficientNet provided competitive accuracy with lower computational cost. These results suggest that both models can serve as effective components in multilingual sign language recognition systems. However, the overall performance also highlights the importance of dataset

diversity, controlled preprocessing, and balanced training samples when working with multiple sign languages. The outcomes of this research offer a clear direction for future work, particularly in expanding the system toward real-time applications and larger multilingual datasets.

7.2 Solutions to the Challenges

There were a number of solutions utilized in the process of the research to ensure that the various challenges that appeared during the development process of the multilingual sign language classification system were solved. Among the initial actions were the efforts made to make the preprocessing pipeline stronger to minimize the differences that exist in the BdSL, ASL, and ISL datasets. Image resizing, normalization, and regular conversion of colors were used as techniques to produce a standardized input of the models. The distribution of classes was also given additional consideration and the datasets were properly balanced to make sure that each class of gestures was represented adequately. This not only enhanced stability of learning but minimized chances of biased predictions.

Mathematical difficulties related to the process of training deep learning modules were addressed through adopting an organized training approach. The training was split into two parts first feature extraction with frozen layers and then fine-tuning where the chosen layers were unfrozen to continue learning. The method minimized the training time and at the same time enabled the models to adjust to the nature of multilingual sign gestures. Early termination and learning rate scheduling were also implemented in order to avoid overfitting and keep the convergence constant. In EfficientNetB0, hyperparameters were optimized to obtain a controlled experiment by successively testing several configurations until the performance was stable.

In an effort to unlock the challenges that come with the incorporation of several sign languages, the data set preparation was geared towards ensuring the equal representation of all languages and the similarity in labeling. This had the advantage of making the model learn general visual patterns and still differentiate language specific differences. The system was in a position to learn all three sign languages effectively because of the balanced and unified approach. All these solutions were part of a better performance of the models and a more reliable sign language classification system.To overcome the difficulties associated with integrating multiple sign languages, the dataset preparation process focused on maintaining equal representation and consistent labeling across all languages. This ensured that the model learned general visual patterns while still distinguishing between language-specific differences. By maintaining a balanced and unified approach, the system was able to learn effectively across all three sign languages. These solutions collectively contributed to improved model performance and a more reliable sign language classification system.

Chapter 8

Future Work & Conclusion

8.1 Conclusion

The invention of this system of classifying multilingual sign languages is a milestone towards making communication more accessible to the deaf and hard-of-hearing society. Through the integration of datasets of Bangla, American, and Indian sign languages and the application of the advanced models of the deep-learning frameworks, including ResNet50 and EfficientNetB0, the project showed that modern machine learning models can successfully identify and classify multiple language hand gestures. The findings indicated a high level of performance in all three sign languages, which established that deep neural networks could learn both shared and language specific visual patterns. This achievement offers a base on the prospective collaboration which will entail the creation of more sophisticated and easy-to-use sign language translationally assistive devices.

During the research process, various limitations were experienced, such as variations in datasets, class imbalance, and constraints in computers. Nevertheless, these concerns were overcome with prudent preprocessings, guided training plans, and fine-tuning of the models and this resulted in stable and accurate model performances. The system not only reached its technical goals but also pointed out the larger opportunity on the use of technology to facilitate inclusive communication. With the growing importance of digital accessibility, such solutions will contribute to the reduction of communication barriers and facilitate the equal access to education, employment opportunities, and the provision of state services.

To sum up, the project has shown that it is possible to have an image-based classification model with reliable accuracy to determine whether the sign language is multilingual. Although the system is centered on the stationary gesture recognition, it offers a solid skeleton on which the further development can be presented, such as the continuous sign recognition, mobile gesture recognition, and something wider references. Through further growth and community cooperation, this piece may have an impact on making the sign language user more accessible and

inclusive to various areas.

8.2 Future Work

Even though the multilingual sign language classification system had positive results, there are various ways on which the work can further be extended in the future. A bigger and more diverse dataset would be one of the most significant changes. The present system is loading on fixed pictures and even though it has been effective in identifying individual gestures, sign language communication usually constitutes sustained movement, switching and facial expression. The addition of video-based datasets or time-varying models (i.e., LSTM or Transformers) may greatly improve the capacity of the system to read entire sign sentences and not to analyze single gestures. This development would bring the system closer to the actual world where fluid communication is necessary.

The second opportunity that might be an avenue of future work is having a lightweight mobile or embedded version of the system. Although presently EfficientNetB0 offers an efficient architecture, further compression of the models can be achieved by adding pruning, quantization, or knowledge distillation to enable the system to be used in real-time on smartphones or low-power machines. That would enable the technology to be accessible to many people including those in rural or underserved communities where they have little access to high-end devices.

Lastly, it would be more important to expand the system to accommodate additional sign languages and regional differences. There exist sign languages varied in different countries, and in some cases even among communities living in the same area. The system would be more inclusive and applicable to other parts of the world if more languages are added to it and the cultural diversity is increased in the set. An addition of facial expression recognition and hand-tracking techniques might enhance accuracy as well, particularly in those signs that need to combine both hand movements and facial expression. Such improvements of the future would aid in making the system more comprehensive and versatile in terms of translating the sign language.

Bibliography

[1] Y. Saleh and G. Issa, "Arabic sign language recognition through deep neural networks fine- tuning," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 16, pp. 71–83, 05 2020.

[2] N. G. Society, "Sign language." National Geographic Society. Online, 2024. Available: `https://education.nationalgeographic.org/resource/sign-language/`, accessed on: Sep. 14, 2025.

[3] S. A. Shaba, S. I. Shroddha, M. J. Hossain, and M. F. Monir, "Neuralgesture communication: Translating one sign language to another sign language using deep learning model and gtts," in *2024 IEEE 99th Vehicular Technology Conference (VTC2024-Spring)*, pp. 1–5, 2024.

[4] S. Das, S. Imtiaz, N. Neom, N. Siddique, and H. Wang, "A hybrid approach for bangla sign language recognition using deep transfer learning model with random forest classifier," *Expert Systems with Applications*, vol. 213, p. 118914, 09 2022.

[5] R. Pranto and S. Siddique, "Real-time bangla sign language translator," 12 2024.

[6] S. K. Akash, D. Chakraborty, M. M. Kaushik, B. S. Babu, and M. S. R. Zishan, "Action recognition based real-time bangla sign language detection and sentence formation," in *2023 3rd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*, pp. 311–315, 2023.

[7] A. Agency, "Hearing-impaired demand sign language institute in bangladesh." Anadolu Agency. Online, 2021. Available: `https://www.aa.com.tr/en/asia-pacific/hearing-impaired-demand-sign-language-institute-in-bangladesh/2371702`, accessed on: Sep. 14, 2025.

[8] M. Hadiuzzaman, M. Ali, A. S. Miah, A. Raj Shafi, and J. Shin, "Baust lipi: A bdsl dataset with deep learning based bangla sign language recognition," in *Proceedings of the 3rd International Conference on Computing Advancements*, ICCA '24, (New York, NY, USA), p. 280–285, Association for Computing Machinery, 2025.

[9] S. M. Rayeed, S. T. Tuba, H. Mahmud, M. H. U. Mazumder, M. S. Hossain, and M. K. Hasan, "Bdsl47: A complete depth-based bangla sign alphabet and digit dataset," 2023.

[10] M. Geislinger, "Asl rgb + depth fingerspelling – spelling it out." https://www.kaggle.com/datasets/mrgeislinger/asl-rgb-depth-fingerspelling-spelling-it-out, 2023. Accessed: 2025-12-08.

[11] d. c. Prathumarikeri, "Indian sign language (isl) dataset." https://www.kaggle.com/datasets/prathumarikeri/indian-sign-language-isl, 2023. Accessed: 2025-12-08.

[12] Google, "Google colaboratory." https://colab.research.google.com/, 2023. Accessed: 2025-09-15.

[13] "Numpy." https://numpy.org/. Version 1.23+, Last accessed: 2025-01-12.

[14] "Pandas." https://pandas.pydata.org/. Version 1.5+, Last accessed: 2025-01-12.

[15] "Tensorflow." https://www.tensorflow.org/. Version 2.12.0, Last accessed: 2025-01-12.

[16] "Keras." https://keras.io/. Version 2.12+, Last accessed: 2025-01-12.

[17] "Pytorch." https://pytorch.org/. Version 2.1, Last accessed: 2025-01-12.

[18] "Opencv." https://opencv.org/. Version 4.7.0, Last accessed: 2025-01-12.

[19] "Matplotlib." https://matplotlib.org/. Version 3.7+, Last accessed: 2025-01-12.

[20] "Scikit-learn." https://scikit-learn.org/. Version 1.2+, Last accessed: 2025-01-12.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *scikit-learn: Machine Learning in Python*, 2011.



Multilingual Sign Language Translation using image-Based classification models for Bangla, American, Indian Languages with ResNet and EfficientNet

By

Sudipto Biswas*

Student ID: **2222391***

Autumn, 2025

The student modified the internship final report as per the recommendation made by his or her academic supervisor and/or panel members during final viva, and the department can use this version for achieving.

Signature of the Supervisor

Saadia Binte Alam, PhD
Assistant Professor and Head
Department of Computer Science & Engineering
School of Engineering, Technology & Sciences
Independent University, Bangladesh