Comparative Analysis of Bangladeshi Sign Language Classification Using Transfer Learning and Fast Fourier Transformation

by

S.M Minoor Karim
21101111
Mahdi Uddin Ahmed
20301469
Himel Banik
19101633
Jawad Ibn Mamoon
20301474
Tawhid Mollah Orpon
21301719

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

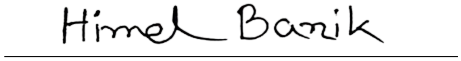
Department of Computer Science and Engineering
School of Data and Sciences
Brac University
January 2023

Declaration

It is hereby declared that

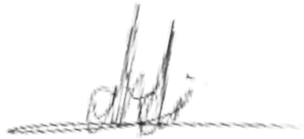
1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:



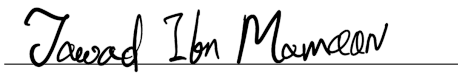
Himel Banik

19101633



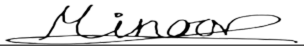
Mahdi Uddin Ahmed

20301469



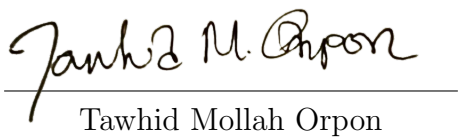
S.M Minoor Karim

21101111



Jawad Ibn Mamoon

20301474



Tawhid Mollah Orpon

21301719

Approval

The thesis/project titled Approach to Bangladeshi Sign Language Detection Using Fast Fourier Transformation" submitted by

1. Mahdi Uddin Ahmed (20301469)
2. S M Minoor Karim (21101111)
3. Jawad Ibn Mamoon (20301474)
4. Himel Banik (19101633)
5. Tawhid Mollah Orpon (21301719)

Of Summer, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on September 18,2023.

Examining Committee:

Supervisor:
(Member)

Farig Yousuf Sadeque
Assistant Professor
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Md Tanzim Reza
Lecturer
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Dr. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Chairperson and Professor
Department of Computer Science and Engineering
Brac University

Abstract

Sign Language, a non-vocalized means of communication via the use of hand signs, gestures and motions to articulately express oneself and communicate with others for differently-abled people, especially hearing and speaking-impaired people. It is the most dominant form of communication in these communities. However, like all forms of communication, this too requires both sides to understand the language, which is the main barrier among many people of the community, due to not only having varied forms of sign language but also very few people outside the community who understands them. Our research aims to design a system using a glove or camera that reads Bangla Sign Language, processes and interprets it efficiently, and provides output in the form of Bangla Language for the other party to understand. We aim to focus on the Pre-processing aspect. Hence we decided to deploy Fast Fourier Transformation to decrease the pre-processing time.

Keywords: Sign Language; Bangla; Detection, Pre-Process; Fast Fourier Transform; interpret.

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.
Secondly, to our advisor Farig Yousuf Sadeque sir and co-advisor Md Tanzim Reza sir for their kind support and advice in our work. They helped us whenever we needed help.
And finally to our parents without their throughout sup-port it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Thoughts behind the Prediction Model

Sign language is an essential means of communication in hearing-impaired communities. Bangladesh houses approximately 2.34 million people with different sorts of disabilities. Bangladesh, with more than 30 lakh hearing-impaired people, lacks adequate opportunities for learning sign language, making it exponentially harder for the community to communicate with others, resulting in their day-to-day life, employment status, and economic status being severely affected. Only 15% of people of these communities get employed[17][16] .
Hence, the main task of this research is to produce an efficient interpreter for people to bridge the gap between communication. The two known methods of sign language detection are Hand Gesture Recognition using Computer vision (CV) and using instrumented gloves[12]. In this research we intend to use both computer vision and gloves for the detection of sign language to facilitate the communication, with gloves for the differently abled person using Bangla Sign Language to communicate and computer vision for people speaking in different Sign Languages to understand them.

1.2 Problem Statement

This research is to address:
"The problem of identifying and analyzing Bangla Sign Language notations using state-of-the-art computer vision and machine learning techniques side by side with the application of U2 net Architecture and Fast Fourier Transform."

With the secondary problem being
"Forming coherent statements with the detected lines using Natural Language Processing."

1.3 Description of the Problem:

Bangladesh often lacks the infrastructure to support its differently abled community. With over 30 lac people being a part of that and only 15% being employed, it is a severe problem. There exist very few opportunities to learn sign language. In our

country, hearing- and speech-impaired individuals do not have the opportunity to study their native tongue. There are no recognized research centers dedicated to advancing and promoting the Bangla sign language. In 1994, the Ministry of Social Welfare and Bangladesh National Federation of the Deaf created a Bangla sign language dictionary. Since then, no meaningful action has been performed other than granting Bangla sign language official status. Despite being the second most widely spoken language, there are little opportunities to study sign language in Bangladesh [10]. With this research we intend to build a system that connects the bridge between the two.

In terms of Detection of Sign language, one of the most common methods of detection is Image processing. Generally real time hand gesture detection coupled with an image processing algorithm has been used in most cases. [18] HMM model, CNN Architecture are two of the most popular models yielding high accuracy. CNN being the newest and most advanced for detection. However, the issue with image processing is quality of image, as most people in Bangladesh live below the poverty line, hence it's reasonable to assume there is a sufficient lack of high quality cameras. Even if that's ensured with the application of colored gloves, the secondary issue lies with the light. We will apply different models to determine the most efficient route to detect Bangla Sign Language using computer vision and prepare a data set for future use. Another approach, as mentioned to SLR using CV is using coloured gloves. In this approach, colored marker hand gloves are worn in the hand. The color of the glove is used to track the movement , position and motion of the hand. Its limitation is that it is also not a natural method of computer human interaction. [12] However the real time detection will only work in well light areas. Hence, the system will be ineffective at night and in dark areas[15]. Which is one of the big issues. As communication is needed in both night and day.

In the context of evaluation scenarios and testing conditions, the accuracy is reasonably high for CV(95% on average) based SLR approach, yet it is not sufficient when a really reliable and resilient system is required. Highly efficient inertial and orientation sensors include accelerometers, magnetometers, and gyroscopes. Unlike vision systems, these devices are unaffected by external circumstances such as illumination and background. These sensors also enable the very simple acquisition of data that are difficult to obtain in vision systems, such as hand shape or forward/backward movement [3]; these parameters include hand shape and forward/backward movement.We will implement different algorithms to read the readings from sensors of gloves, along with testing a multitude of sensors to find out the optimal route to detect sign language using gloves. Fuzzy Min-Max to recognize the input from a Data-glove yielded 85% accuracy in detecting Korean Sign Language KSL.[18] HMM, PaHMM had also shown huge accuracy coupled with accelerometer equipped gloves.[3] However, the issue with such a system is requirement of sophisticated systems, often costly followed by less accuracy then its computer vision counterpart.

Another issue is, after detection of the Sign Language, coherent sentences need to be formed. Bangla language has many words which can be used for different meanings depending on the sentence and context. Hence, as a secondary focus, we will implement an NLP algorithm to approach this issue. The main focus of this research is to detect sign language.

In conclusion, there is a severe lack of infrastructure to bridge the gap between hearing or speaking impaired communities with other communities, with very less

work done based around Bangla Sign Language Recognition (BdSLR). Using the aforementioned data the research intends to find a suitable solution and data sets for future research.

1.4 Research Objective

The study intends to create a Bangla Sign Language Detection system applying both Computer vision and Instrumented Gloves that can accurately detect Bangla Sign Language using Hand gesture movement in real time, process the input and with application of NLP provide a text/Speech output to facilitate communication between both parties. Using different hand gesture detection algorithms, NLP, Sensor based gloves, we will prepare a dataset and compare the result of the experiment and determine the optimal route to detect Bangla Sign Language. The following are the study goals.

- To gain an understanding of Sign Language Detection and how it works.
- To learn about Computer vision and Recognition algorithms.
- To familiarize with Hand Gesture Detection algorithms.
- To attain an understanding of data collection and processing
- To gain an understanding of Hand gestures Detection using sensor gloves.
- To compare different algorithms with respect to Bangla Sign Language.
- To Suggest an efficient way to detect Bangla Sign Language.

Chapter 2

Litrature Review

This article will show how different image processing algorithms have been used to identify different forms of sign language across the world. We shall study the effectiveness of different methods implemented and the resulting output. Along with that we shall study the process of sign language communication and ways to use sensors to detect the sign language and discuss different software and hardware architectures.
Sign language is a common form of communication for differently abled people struggling with hearing or speech related problems. Different nations have their own form of sign language. Issue is a lot of people do not understand the language hence making the lives of vulnerable communities harder. With the application of this research a process will be created where people can communicate easily without the need to understand sign language. The research will mostly help speech impaired people of Bangladesh.

2.1 Image Processing:

The primary issue with Sign Language Detection using computer vision is light. Most algorithms struggle to detect hands properly in less lit areas. Existing sign language recognition algorithms have a low identification rate in environments with a cluttered background and variable illumination, environmental noise, training situations, and a high computing cost[13]. To solve this we will use a Hand Cascade classifier and Skin color segmentation paired with a gaussian filter to reduce noise and accurately detect hand signs. For the image processing algorithm different algorithms such as HMM or CNN architecture which had ensured above 90% accuracy. They demonstrated a dataset of how each individual algorithm yields a different rate of accuracy to the same dataset. We also focused on studying how these algorithms have been used to identify hand gestures in other forms of research and how they performed.

2.2 Hand Detection:

The first step of Sign Language Processing with Computer Vision is Hand detection. Following which different architecture is used to process and identify the sign. One of the methods is Edge detection using the HSV model.

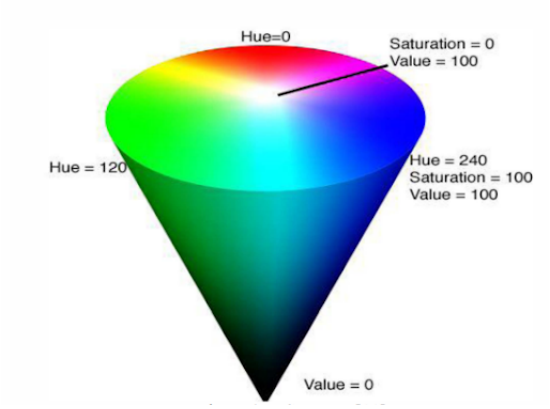


Figure 2.1: HSV Color model Cone

In the case of the process used by J. Knowar [8], we see initially they took a coloured image of the hand. Using matlab functions they converted the RGB Color to grayscale. Using the HSV model we can map the hue and saturation of skin pixels. Using that a hand shape is formed using the HSV model. It uses the canny algorithm to detect the edges. In this process we initially remove the noise using a gaussian filter, followed by using 2D spatial gaussian measurement to find the edge strength at each point of the image. Finally the magnitude or strength , of gradient is then approximated using $\|G\| = \|Gx\| + \|Gy\|$.
X, Y value gives the edge direction. Next the values are mapped in a direction that can be traced in an image. Then after tackling the streaking issue, binary output of the image is prepared for structuring the edges, generally the process is called morphological operation

Figure 2.2: Binary Image Output for Edge

2.2.1 U^2-Net Model:

However, upon further research, we discovered A model called U^2-Net was introduced in a paper by Xiaojuan Qi, Jianping Shi, Xiaogang Wang and Hongsheng Li from the Nanyang Technological University. The model, published in the proceedings of the IEEE Conference on Computer Vision and Pattern Recognition in 2020, is a deep neural network architecture for salient object detection. The U^2-Net uses a nested U-structure to effectively capture both local and global context information in an image and outperforms existing state-of-the-art methods on several benchmark datasets for salient object detection.

However, U^2-Net Model is not specifically designed for background removal, it can be used for this task by identifying and segmenting salient objects in an image and treating the remaining pixels as the background. This approach can be effective for some types of images, but may not work well in all cases, especially in images with complex backgrounds or similar color to the object. It would be interesting to see how the U^2-Net model performs on background removal tasks and compare it with other existing methods designed specifically for this task.[20]
A study was conducted on highly accurate dichotomous image segmentation using the U^2-Net model. The U^2-Net is a deep neural network architecture that was designed for salient object detection, which aims to identify and highlight the most "salient" or "attention-grabbing" regions in an image. The U^2-Net model is based on a nested U-structure, which is designed to effectively capture both local and global context information in an image. The authors of the study used this model to perform dichotomous image segmentation, which is the process of separating an image into two distinct classes.
The study found that the U^2-Net model was able to achieve high accuracy in dichotomous image segmentation tasks when compared to other existing methods. The U^2-Net's ability to capture both local and global context information made it well suited for this task, as it was able to effectively identify and separate the two classes in the image. The study also highlighted that the U^2-Net's performance was consistent across a wide range of images, and it was able to maintain high accuracy even when dealing with images that had complex backgrounds or similar color to the object.

2.3 Image processing and classification Architecture:

Numerous techniques for sign language recognition have been developed to date. Kulkarni utilized the HSV color model for segmentation as well as the Hough algorithm and histogram technique to extract features. Three-layer feed forward neural networks are utilized for gesture classification. Wysoski et al. introduced rotation-invariant postures using a boundary histogram. The HMM was utilized by Hyeon-Kyu Lee and Jin H. Kim for categorization purposes (Hidden Markov Model). Stergiopoulou approximated hand shape morphology using YCbCr color space and a Self-Growing and Self-Organized Neural Gas (SGONG) network [8].

Faster R-CNN, a method derived by refining Fast R-CNN, which is itself derived by enhancing R-CNN, is used to detect Bangladeshi Sign Language, as demonstrated in [5]. Input frames are divided into anchor maps based on CNN. Through the process of RPN, these anchor points are established. The bounding boxes of each anchor point are analyzed and processed, labeling them as objects or backgrounds. This process is done with the use of RoI Pooling. After compiling all labeled data, it is further refined based on previously defined criteria. The final part of the process is completed using R-CNN Based on this refined data, the hand movements and gestures are identified and translated.

When looking at the limitations of the paper in [5], it can be seen that there were

many major shortcomings of the system developed there. To start with, there was an issue with recognition, where letters were mixed up due to subtle similarities. Another limitation was that all test data was collected in controlled environments and from a limited subject pool. As it stands, the system is also unable to properly establish sentences or phrases, as it processes every individual sign separately without sequencing.

2.3.1 Fast Fourier transform:

In the study "Arabic Sign Language recognition using FFT and PCNN", the authors use the Fast Fourier Transform (FFT) as a key component of their method for recognizing Arabic Sign Language (ArSL).

The FFT is a mathematical algorithm that is commonly used in signal processing to transform a time-domain signal into the frequency domain. It decomposes a signal into its individual frequency components, allowing for a better understanding of the underlying frequencies present in the signal. In this study, the authors use the FFT to analyze the time series of the firing status of neurons obtained from PCNN. By transforming the firing status of neurons into the frequency domain, the FFT allows the authors to extract features that are related to the frequency of the firing. They use a maximum of 11 FFT descriptors as features for their recognition system. The authors claim that the use of FFT for feature extraction improved the recognition accuracy compared to previous methods, as it allows them to extract more meaningful and discriminative features from the signals, making the system more robust to variations in signer's style.

It's worth noting that FFT is a widely used technique in signal processing, speech and image processing, and it's not limited to just recognizing ArSL, it could be used to extract features from other signals such as speech, audio, and other time series data.[22]

Chapter 3

Description of the model and data

3.1 Methodology:

- To create a custom dataset. We create a custom Bangla Sign language dataset using Ishara Lipi convention. We recorded different signs representing different alphabet to create our dataset. Different skin tone was ensued to create a variety. After the recording we extracted the images and labeled them. We then start pre processing.

- To apply U2-Net background remover to the dataset. U^2-Net is a deep learning model that can be used to remove the background of an image, leaving only the foreground objects. This can help to improve the performance of the model by reducing the amount of noise and distraction in the images.

- To apply FFT compression to the dataset. FFT (Fast Fourier Transform) compression is a technique that can be used to reduce the amount of data in an image while maintaining its important features. This can help to speed up the processing time of the model, as well as improve its performance by reducing the amount of noise and distraction in the images.

- To apply a FFT high-pass filter for edge detection. FFT high-pass filter is a technique that can be used to enhance the edges of an image, which can be useful for detecting signs and gestures in the dataset. This can help to improve the performance of the model by making it more sensitive to the important features of the images.

- To Label the pre processed images. We have to label all the images and divide them into a 80-20 split for training and testing purpose. We do it with both the raw data and pre processed data.

- All of this preprocessing is done to decrease the parameters for the Models. We aim to decrease the parameters and ensure maximum accuracy while decreasing training and detecting time. To pass the pre-processed dataset to different models like VG19,Resnet50 and Mobilenet for Bangla Sign Language detection. VG19 is a convolutional neural network (CNN) model that can be used for image classification, while Mobilenet is a lightweight CNN model that can be used for real-time object detection.

- To train the models using the pre-processed dataset. This will involve using a large number of images to train the model, and adjusting the model's parameters to optimize its performance.
- To test the models on a separate validation dataset. This will involve using a smaller number of images to test the model, and evaluating its performance by comparing its predictions to the actual labels of the images.
- To train the raw data. We use the initial data we labeled before pre processing and note down the result. We then compare the data before and after pre processing to see if there is a major impact in models performance.
- Once the model is trained and validated, use the model for real-time detection of Bangla Sign Language. This can be done by using a webcam or other device to capture live video, and then using the trained model to classify the signs and gestures in the video in real-time.

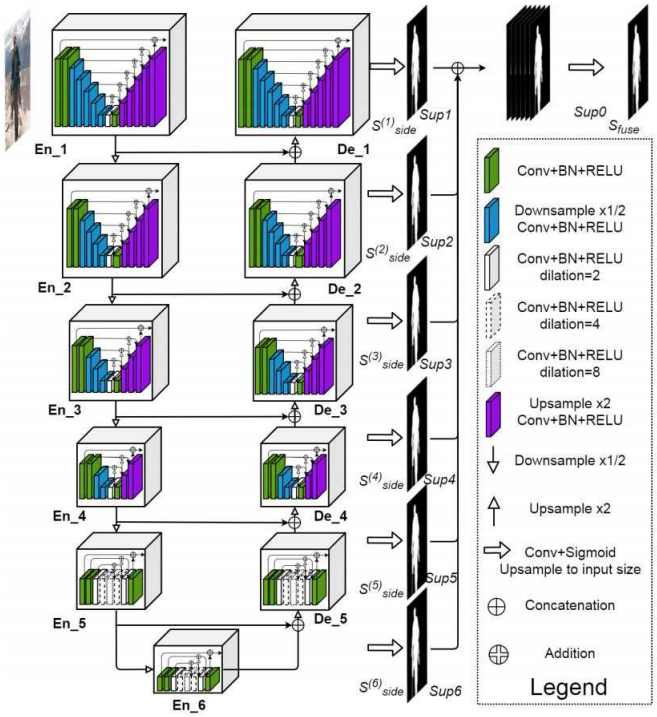


Figure 3.1: U^2 Net Architecture

3.2 Image Pre Processing:

The pre-processing can be divided into 3 steps. Background removing, Gray Scaling, and finally Image Compression using FFT.

3.2.1 Background removal with U^2-Net:

U^2-Net is trained on a dataset of images with corresponding binary masks that indicate the background and foreground regions. The network learns to map the input

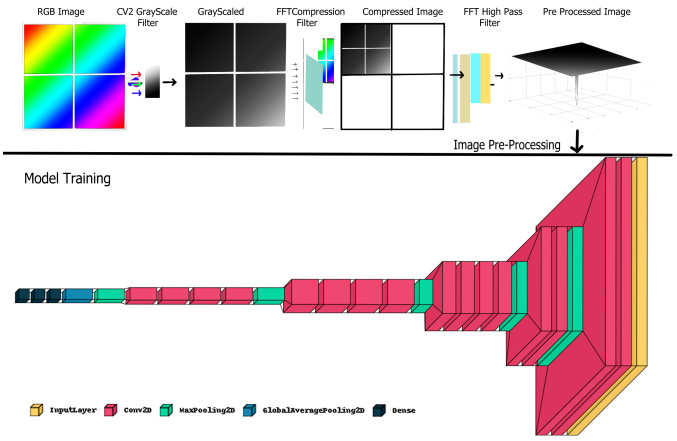


Figure 3.2: Process flow diagram

image to the corresponding binary mask through the encoder-decoder architecture and the squeeze-and-excitation module.The encoder part of the network is used to extract features from the input image and the decoder part of the network is used to upsample the feature maps and combine them with corresponding feature maps from the encoder to refine the segmentation results. The squeeze-and-excitation

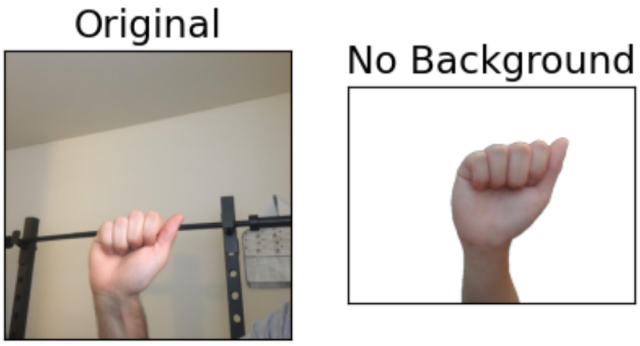


Figure 3.3: U^2 Net output

module allows the network to focus on important features in the image by generating a set of weights for each channel, which are then multiplied with the feature maps to produce the output of the SE module. Finally, the output layer produces the final binary mask that can be used to remove the background from the original image.

3.2.2 Gray Scaling the Image:

Grayscaling an image before applying the FFT can improve the frequency analysis of the image, as it reduces the computational complexity and noise. The FFT algorithm is used to analyze the frequency content of an image, by decomposing it into its sine and cosine components. When an image is in color, it has 3 color channels (R, G, B) which contain different frequency information, but gray scaling an image merges all the color channels into one, providing a more comprehensive frequency representation of the image.

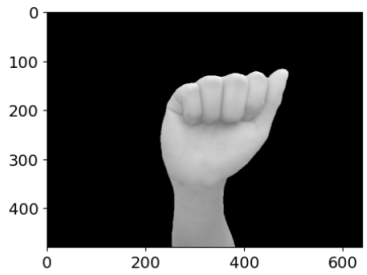


Figure 3.4: Gray scaled image

Additionally, grayscaling removes noise in the form of color artifacts, which can interfere with the frequency information in the FFT, resulting in a clearer and more accurate frequency analysis of the image. This can be done using the cvtColor() function in CV2, which takes in the original image and the desired color space conversion as inputs. For grayscaling, the desired color space conversion is cv2.COLOR_BGR2GRAY.

3.2.3 FFT Compression:

FFT (Fast Fourier Transform) image compression involves representing an image in the frequency domain, rather than the spatial domain. A grayscale image can be represented as a 2D array of pixels, where each pixel is a single value representing the intensity of that point in the image.
The FFT of an image is calculated by taking the 2D discrete Fourier transform (DFT) of the image. The formula for the DFT of a grayscale image is given by:

$$F(u,v) = \sum\sum f(x,y) * e^{(-2\pi i((\frac{ux}{M})+(\frac{vy}{N}))}$$

where, f(x,y) is the pixel value at position (x,y) in the original image
F(u,v) is the pixel value at position (u,v) in the frequency domain
M and N are the number of rows and columns in the original image
i is the imaginary unit

Once the FFT of the image is obtained, it can be compressed by keeping only a certain percentage of the data (e.g. 10%) and discarding the rest. The most significant values in the frequency domain are typically the low-frequency values, as these contain the most important information about the overall structure of the image. The compression works by taking advantage of the fact that the majority of the information in an image is typically concentrated in the low-frequency values of the image's frequency representation. In the FFT of an image, low-frequency values are located at the center of the frequency spectrum, while high-frequency values are located at the edges.

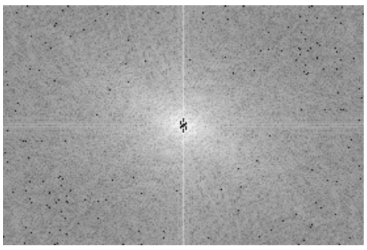


Figure 3.5: Image Frequency Domain

3.3 Model Description:

3.3.1 ResNet50 and ResNet101:

In the study "Deep Residual Learning for Image Recognition" by He et al., presented a new architecture for Convolutional Neural Networks (CNNs) known as "Residual Networks" (ResNet) that aims to overcome the problem of diminishing gradients in deep networks. The authors proposed adding a connection, called "residual connection," between input and output of the same layer and using them to learn residual functions. The residual functions are the differences between the desired mapping and the initial representation. This approach allows the network to learn more efficiently even when it has a large number of layers.

The authors demonstrated that ResNet architecture outperforms previous state-of-the-art CNNs on several image recognition benchmarks, including the ImageNet dataset. They also showed that the network can be easily scaled to hundreds or even thousands of layers and still achieve good performance. This study emphasizes the importance of addressing the problem of diminishing gradients in deep networks and suggests that by using complex CNN architectures such as ResNet, deep learning models can achieve better results and be more robust to variations in the dataset.

3.3.2 MobileNetV2

In "MobileNetV2: Inverted Residuals and Linear Bottlenecks" by Sandler et al., the authors presented a new version of the MobileNet architecture that aims to improve the computational efficiency while maintaining high accuracy. The architecture of MobileNetV2 is based on the concept of using "inverted residuals" and "linear bottlenecks" to improve the efficiency of the network.

Inverted residuals are a type of residual connection that use a linear bottleneck to reduce the number of channels in the input before being passed through a depthwise convolutional layer. This can be represented mathematically as:

$$y = H(x) + F(x, W)$$

where y is the output, x is the input, H(x) is the shortcut or identity mapping and F(x, W) is the residual function that is learned by the network.

Linear bottlenecks are used to increase the capacity of the network without increasing the number of parameters. This can be represented mathematically as:

$$y = F(x, W)$$

where y is the output, x is the input, and F(x, W) is the function learned by the network.

The authors demonstrated that MobileNetV2 can achieve high accuracy on several image recognition benchmarks and it can be easily scaled to different resource constraints. This study emphasizes the importance of addressing the trade-off between accuracy and computational efficiency in deep learning models and suggests that using complex architectures such as MobileNetV2 can improve the efficiency of the network while maintaining high accuracy.

3.3.3 VGG19

In the study "Very Deep Convolutional Networks for Large-Scale Image Recognition" by Simonyan and Zisserman, a new architecture of Convolutional Neural Network (CNN) named VGG19 was presented which is known for its deep architecture and large number of parameters. The authors used multiple small convolutional filters, such as 3x3, in a very deep network, with a total of 19 layers, 16 of which are convolutional layers and 3 are fully connected layers. They argue that using multiple small filters allows the network to learn more fine-grained features and achieve better performance.

The authors showed that VGG19 outperformed other state-of-the-art networks on several image recognition benchmarks, including ImageNet dataset. They also demonstrated that the network can be easily trained on large datasets and still achieve good performance. This study highlights the effectiveness of using deep architectures and small filters in CNNs for image recognition tasks. It suggests that by using a complex CNN architecture like VGG19, deep learning models can achieve better results and more robust to variations in the dataset.

3.4 Dataset Description:

3.4.1 Dataset Description:

The dataset used to train our proposed model was a representation of the American Sign Language(ASL) which has images in different lighting environments and different backgrounds. The dataset contains a total of 29 classes which include the alphabets from A-Z and furthermore has SPACE, DELETE, NOTHING to be represented for use in text transcription. The dataset contains a total of 223 thousand images. The following image represents a sample of different classes in the dataset. We used this dataset to establish our proof of concept while creating our custom dataset.

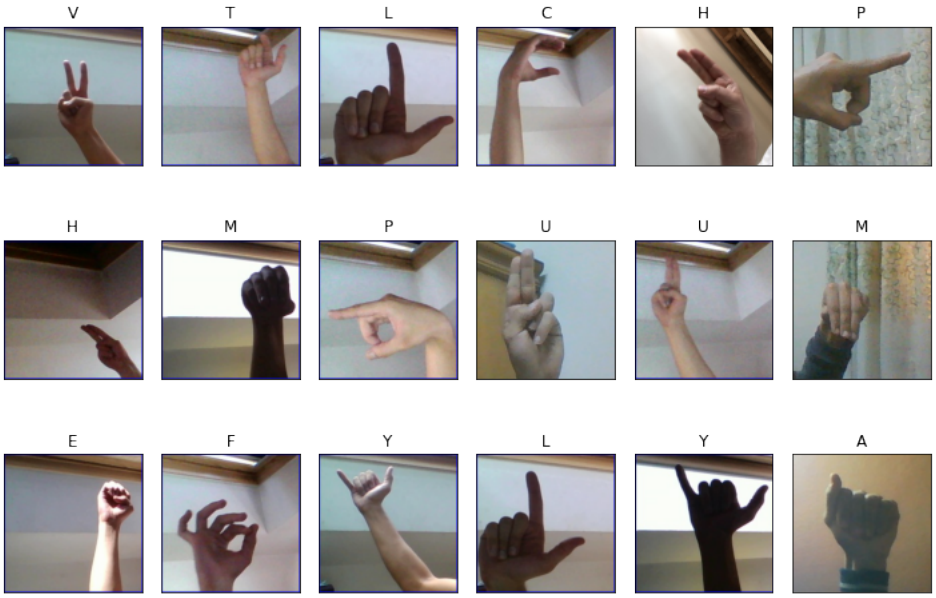


Figure 3.6: Dataset representing American Sign Language

3.4.2 Our Custom Dataset Description:

The dataset used to train our proposed model is custom dataset created by us. There are 10 thousand labelled images. Each alphabet has over 200 images. We created Ishara Lipi Bangla signs for each Bangla alphabet. We recorded around 10 second video each alphabet, extracted them and labelled them. After that we pre processed them and labelled them again.
Each image was 640x640 pixel size. They are RGB images with background as material present.

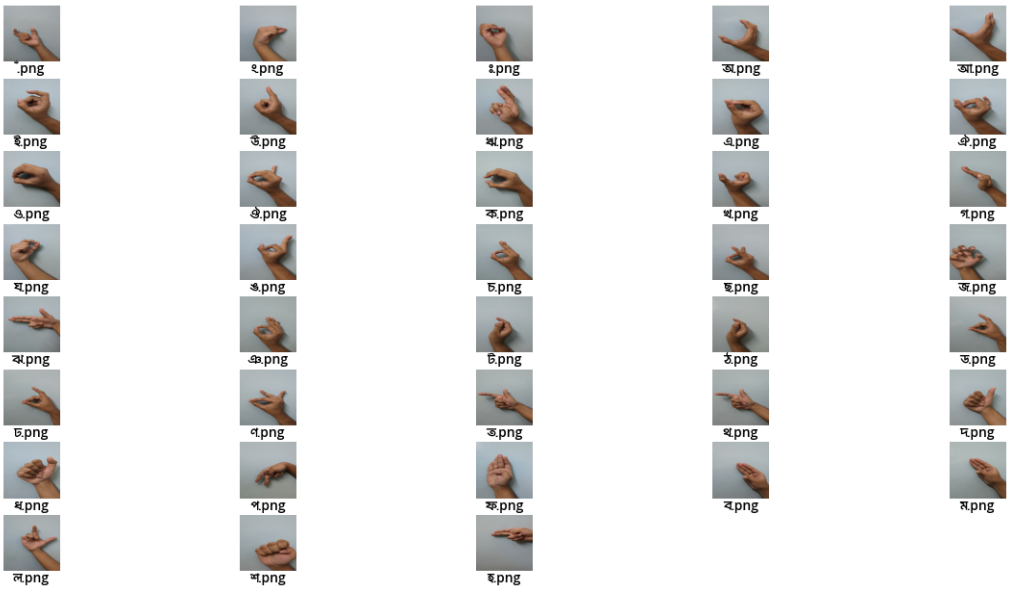


Figure 3.7: Dataset representing Bangla Sign Language

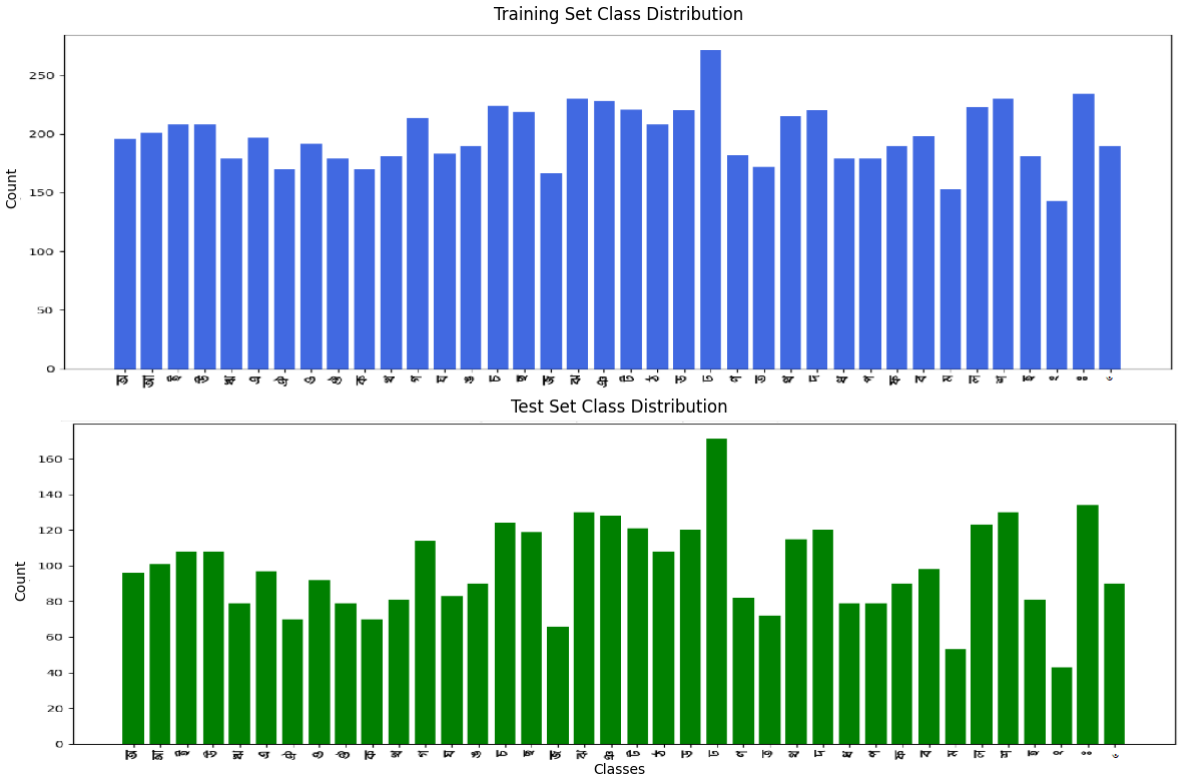


Figure 3.8: Class diagram with Train Test Split

3.4.3 Pre-Processed Dataset:

Here each image is 213x213 pixel in size. The background has been removed using U-2Net algorithm without harming the main focus of the dataset, human hands.

Figure 3.9: Background removed image using u2net

The images have been compressed to 50% by using Fast Fourier Transform, that is we have dropped 50% data from the images.

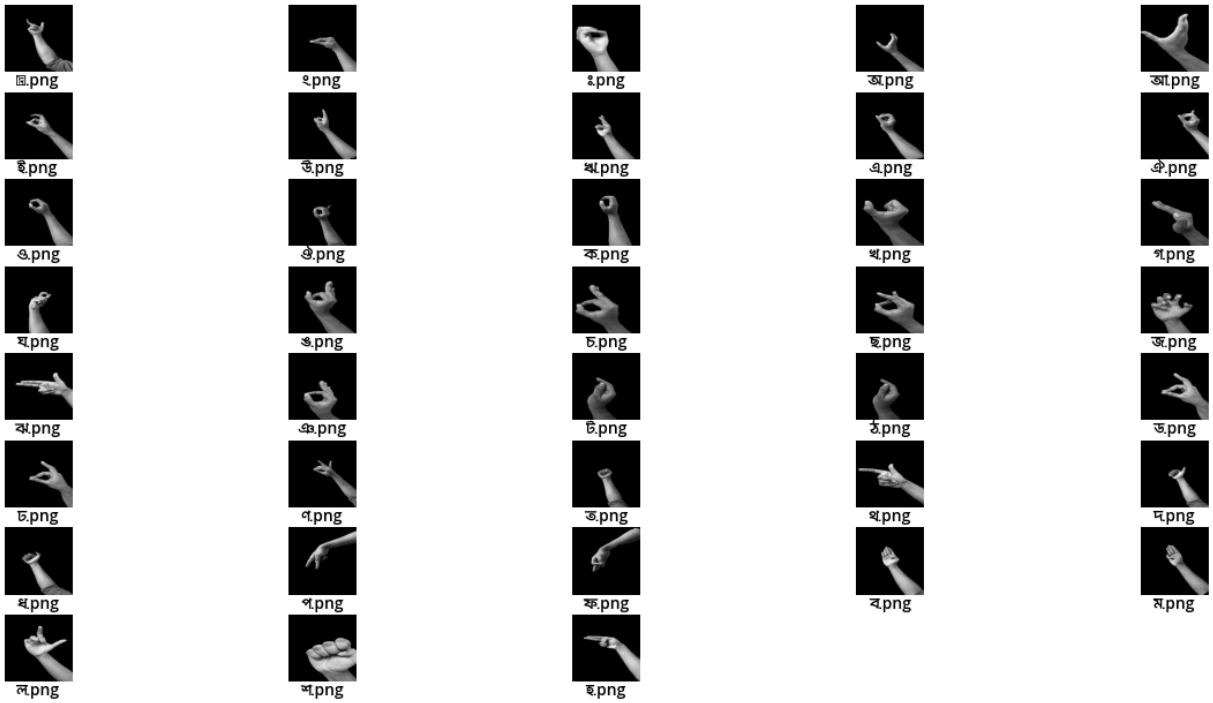


Figure 3.10: Pre processed Images

Chapter 4

Preliminary analysis and Conclusion

4.1 Image Pre processing:

The pre-processing resulted in a severe decrease in parameters. Following graph shows the change in parameters for model training of an image and dynamic range of image at different stages of pre-processing.

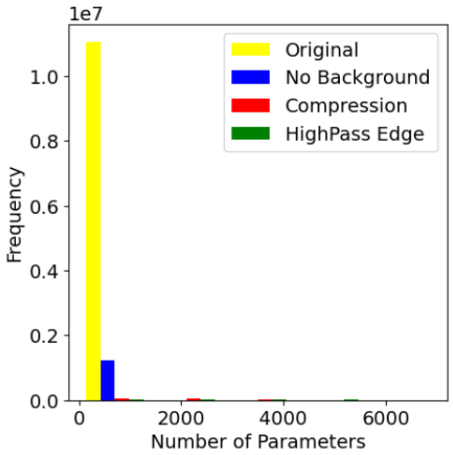


Figure 4.1: Parameters Decrease at different stages

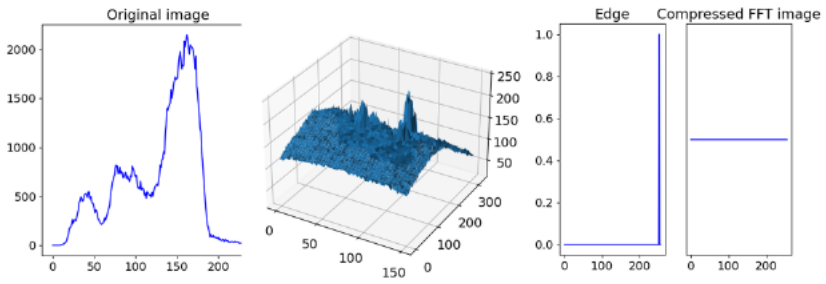


Figure 4.2: Change of Dynamic Range at different stages

4.2 Model Training Result:

We have used the above mentioned dataset on different models to see the accuracy of sign language detection. Following models showed a high amount of accuracy

Model	Test loss	Accuracy
ResNet50	0.11408	96.91%
MobileNet	0.11784	96.30%
VGG19	0.11133	97.24%

Table 4.1: Different model training outcome

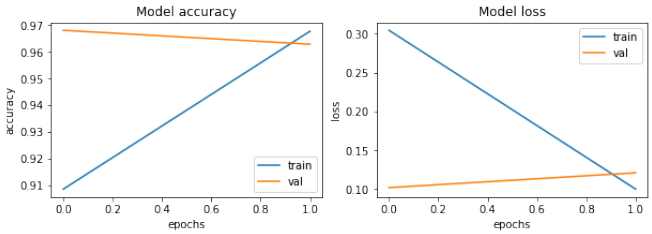


Figure 4.3: MobileNet Accuracy graph

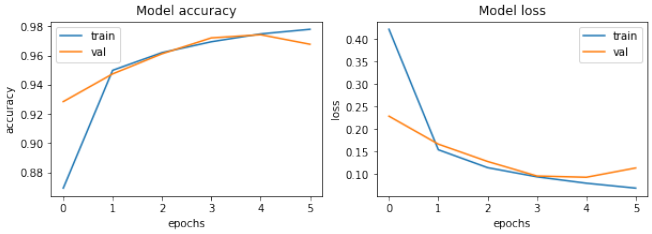


Figure 4.4: ResNet50 Accuracy Graph

4.3 Model Training Result:

Based on that we have implemented Restnet50 and VGG19 for our dataset. We ran it once for regular dataset and then for pre processed dataset. We have noticed two things.

- The pre processed data took very less time, as illustrated in the table below
- To accuracy is unaffected as well as loss.

This signifies we achieved same result within less time and a huge decrease in computational power. We converted the image from 3 channel RGB image to gray scale. With background removed and image compressed to 50% using Fast fourier Transform, we achieved high accuracy within short time.

Regular Dataset				
Model	Epoch	Test loss	Accuracy	Time/ step
ResNet50	30	0.1055	93.48%	901ms
SSD Mobilenet V2	40	0.0961	93.98%	698ms
Inception V3	30	0.1377	93.81%	1s
Pre Processed Dataset				
Model	epoch	Test loss	Accuracy	Time/step
ResNet50	30	0.1392	94.30%	120ms
SSD Mobilenet V2	20	0.0955	94.25%	96ms
Inception V3	30	0.0879	94.14%	233ms
VGG 19	30	0.1220	93.21%	483ms

From the table it is evident that we reduced time for every step in model training. Here we showed how much time it takes for every step in a single epoch. We set the batch size to 32. For regular dataset, VGG19 we had to set batch size to 8 due to GPU limitations. We utilized Google Colabs T4 GPU with 15GB GPU RAM.

4.3.1 Observation:

The VGG19 model couldn't work with batch 32 in those conditions with regular dataset however, with Pre processed dataset we observed it could work with 32 batch size.

We also observe minimal change in loss and accuracy. hence we can conclude we achieved high accuracy without loss of quality with a relatively small dataset and low computational power with our pre processing approach.

4.3.2 Graphical Analysis

Here in the following graph we can see the output of our results compared. We showed Accuracy graph, loss graph, Confusion matrix. In the Second confusion matrix each number represents a Bangali alphabet.

4.3.3 Resnet50

Regular Dataset

Resnet50

Regular Dataset

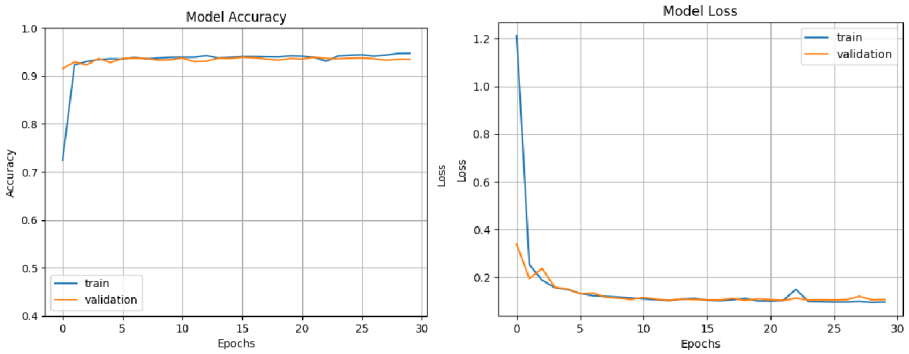


Figure 4.5: ResNet50 Accuracy Graph

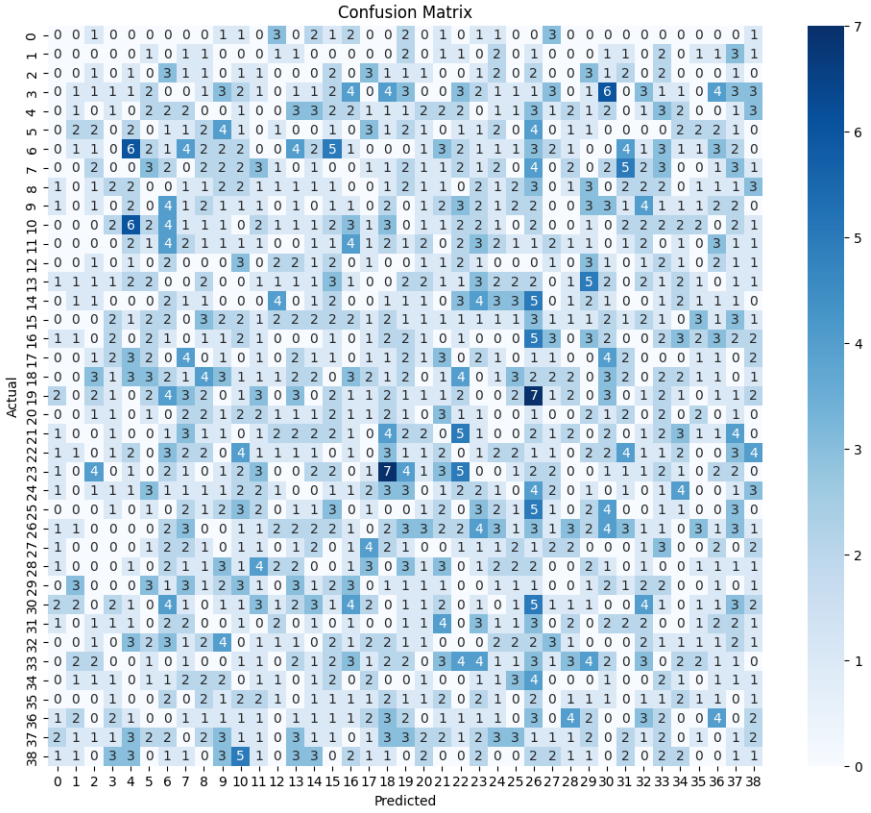


Figure 4.6: ResNet50 Confusion Matrix

Pre-processed

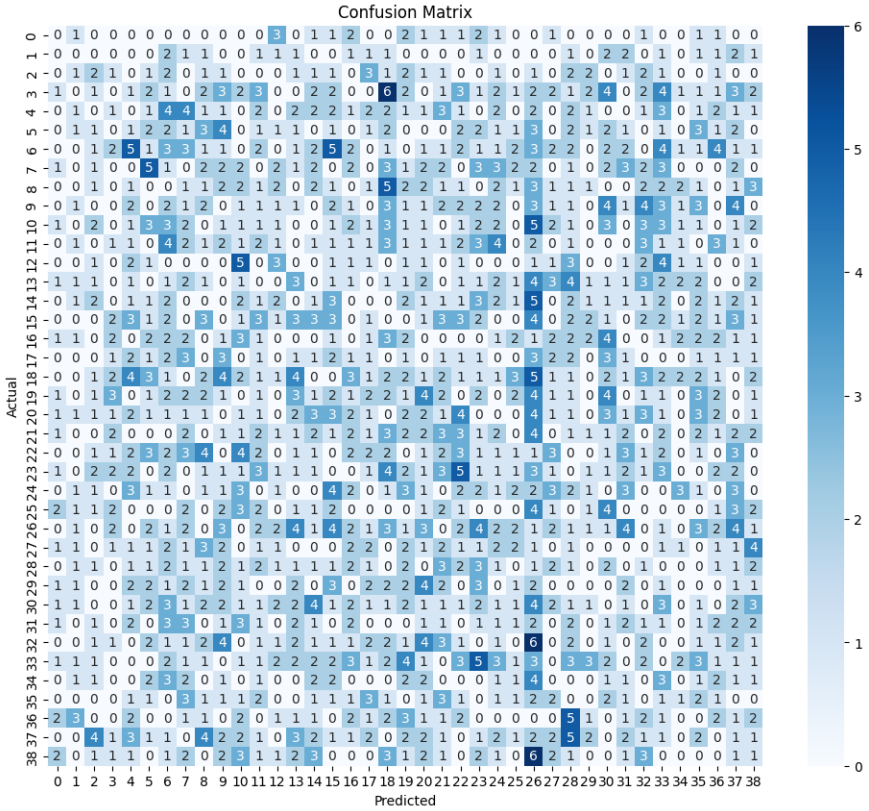


Figure 4.7: ResNet50 Pre processed Confusion Matrix

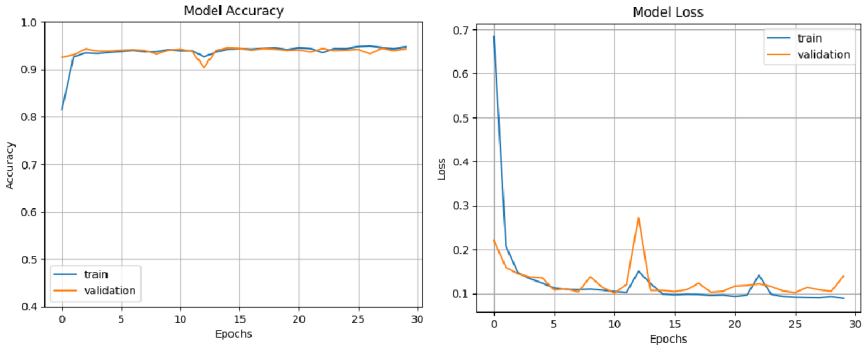


Figure 4.8: ResNet50 Pre processed Accuracy Graph

SSD Mobilenet V2

Regular Dataset Pre-processed

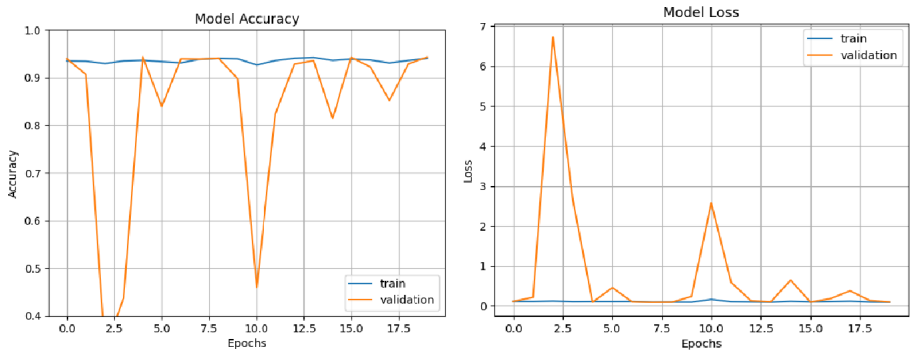


Figure 4.9: Mobilenet V2 Accuracy Graph

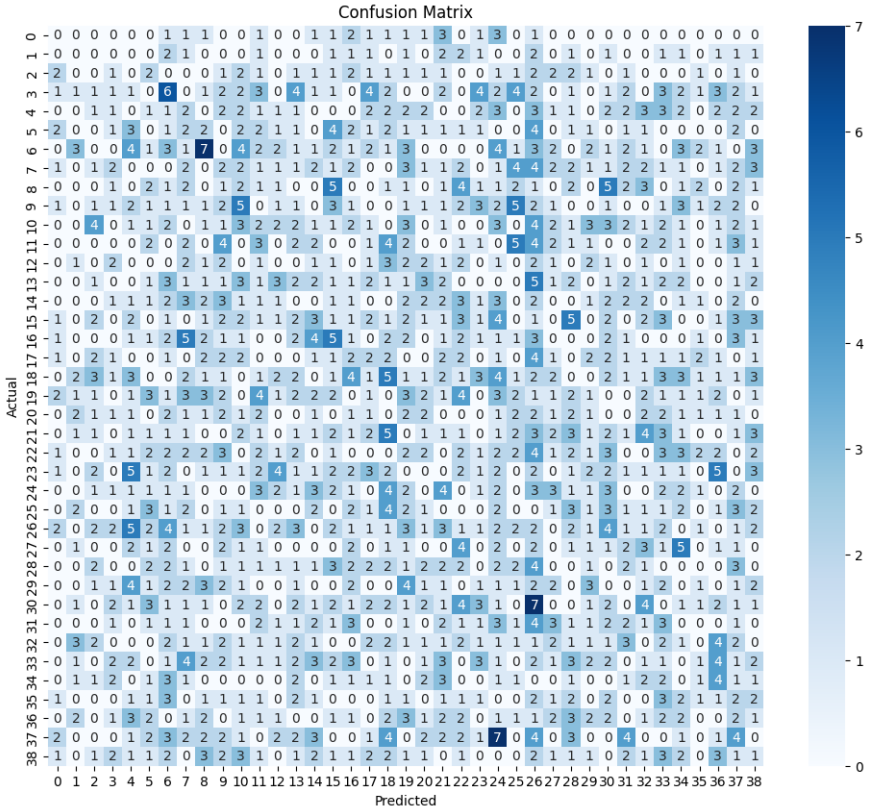


Figure 4.10: Mobilenet V2 Confusion Matrix

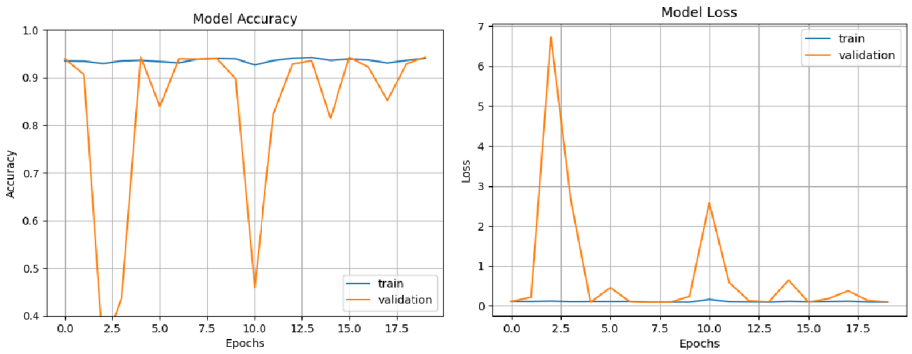


Figure 4.11: Mobilenet V2 Pre Processed Accuracy Graph

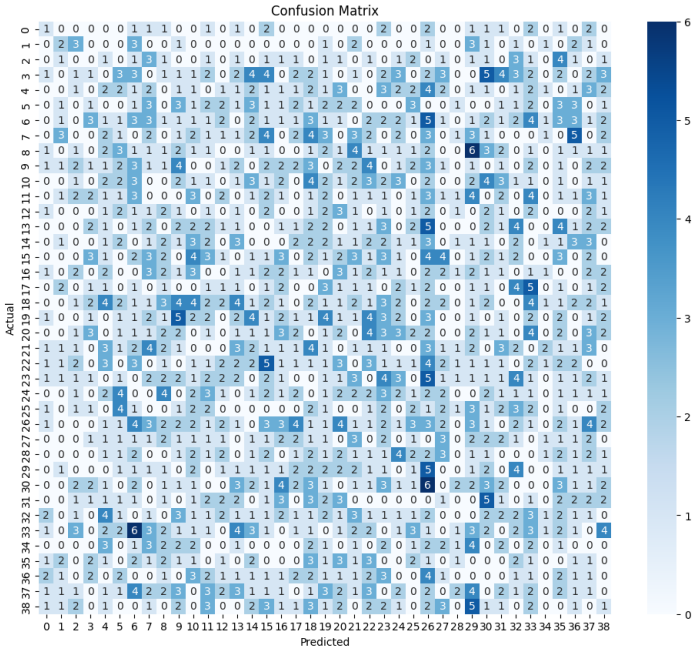


Figure 4.12: Mobilenet V2 Pre Processed Confusion Matrix

Inception V3

Pre-processed

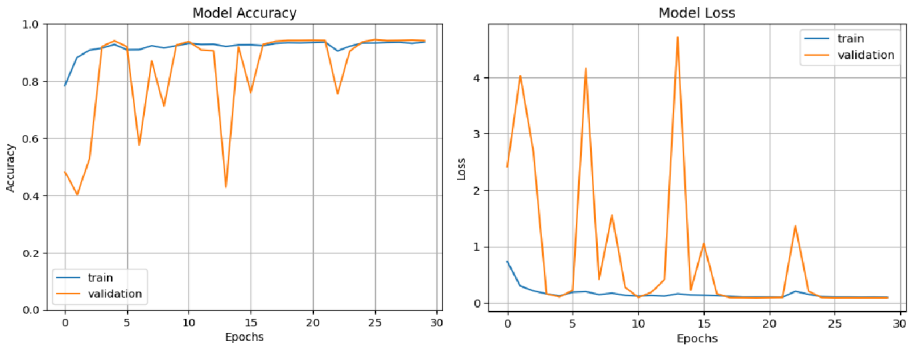


Figure 4.13: Inception V3 Pre Processed Accuracy Graph

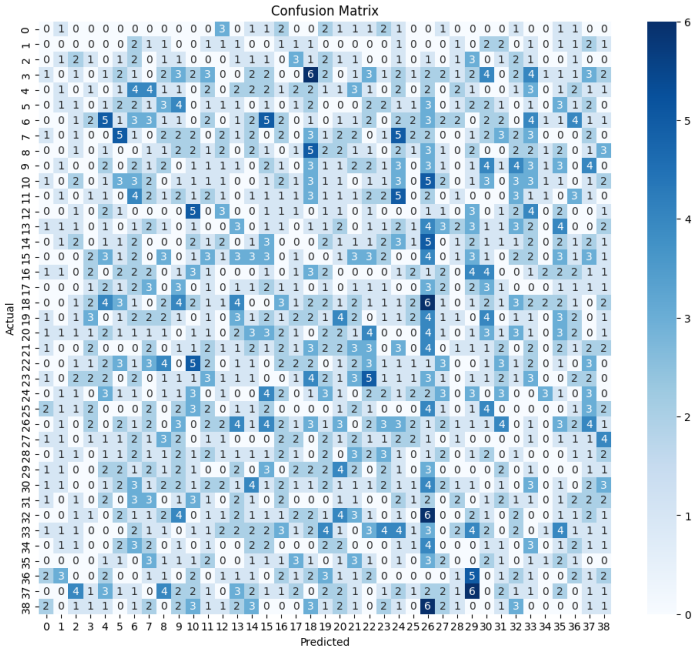


Figure 4.14: Inception V3 Pre Processed Confusion Matrix

.

VGG 19

Pre-processed

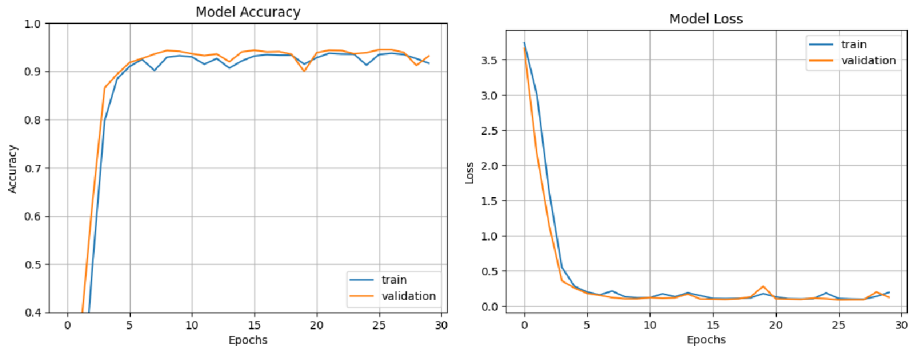


Figure 4.15: VGG19 Pre Processed Accuracy Graph

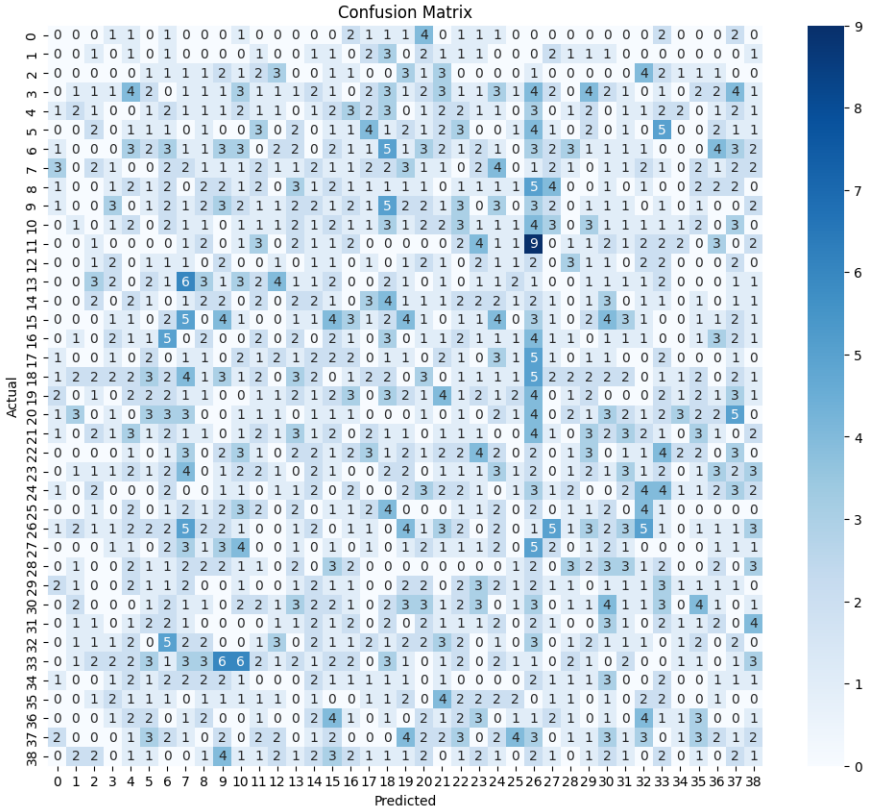


Figure 4.16: VGG19 Pre Processed Confusion Matrix

Chapter 5

Conclusion

5.1 Conclusion:

With this research we expect to help the differently abled community and researchers aiming for the same by finding an optimal route to identify BdSL and to develop a pre-processing technique and an architecture consisting of the most optimal model capable of accomplishing that. The comparative research aims to help multiple future works in the field of BdSLR and in turn help mitigate the hardships of hearing and speech impaired communities of Bangladesh.

Reference

1. Galka, J., Masior, M., Zaborski, M., Barczewska, K. (2016). Inertial Motion Sensing Glove for Sign Language Gesture Acquisition and Recognition. IEEE Sensors Journal, 16(16), 6310–6316.
2. Hoque, O. B., Jubair, M. I., Islam, M. S., Akash, A.-F., Paulson, A. S. (2018). Real Time Bangladeshi Sign Language Detection using Faster R-CNN. 2018 International Conference on Innovation in Engineering and Technology (ICIET).
3. Konwar, A. S., Borah, B. S., Tuithung, C. T. (2014). An American Sign Language detection system using HSV color model and edge detection. 2014 International Conference on Communication and Signal Processing.
4. Moniruzzaman, Md. "Sign Language and Violation of Human Rights in Bangladesh." Bangladesh Journal of Legal Studies, 16 Dec. 2015, bdjls.org/sign-language-and-violation-of-human-rights-in-bangladesh.
5. Pratibha Pandey,Vinay Jain, Hand Gesture Recognition for Sign Language Recognition: A Review. International Journal of Science, Engineering and Technology Research (IJSETR)Volume 4, Issue 3, March 2015
6. Rahaman, Muhammad Jasim, Mahmood Ali, Md Hasanuzzaman,. (2017). A Real-Time Appearance-Based Bengali Alphabet And Numeral Signs Recognition System. 19-26.
7. Shukor, A. Z., Miskon, M. F., Jamaluddin, M. H., Ali Ibrahim, F. bin, Asyraf, M. F., Bahar, M. B. bin. (2015). A New Data Glove Approach for Malaysian Sign Language Detection. Procedia Computer Science, 76,
8. Sakib, SM Najmus. "Hearing-Impaired Demand Sign Language Institute in Bangladesh." Hearing-Impaired Demand Sign Language Institute in Bangladesh, 22 Sept. 2021, www.aa.com.tr/en/asia-pacific/hearing-impaired-demand-sign-language-institute-in-bangladesh/2371702.
9. Siddiqua, Fayeka. "The Silent Conversation — The Daily Star." The Daily Star, 21 Aug. 2014, www.thedailystar.net/the-silent-conversation-37929.
10. Suharjito, Suharjito Ariesta, Meita Wiryana, Fanny Kusuma Negara, I Gede Putra. (2018). A Survey of Hand Gesture Recognition Methods in Sign Language Recognition. Pertanika Journal of Science and Technology. 26. 1659-1675.
11. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O. R., Jagersand, M. (2020). U2-Net: Going deeper with nested U-structure for salient object detection. Pattern Recognition, 106, 107404. https://doi.org/10.1016/j.patcog.2020.107404
12. Qin, X., Dai, H., Hu, X., Fan, D.-P., Shao, L., Van Gool, L. (2022). Highly Accurate Dichotomous Image Segmentation. ArXiv:2203.03041 [Cs]. https://arxiv.org/abs/2203.03041
13. Sidig, A. addin I., Luqman, H., Mahmoud, S. A. (2017). Transform-based Arabic sign language recognition. Procedia Computer Science, 117, 2–9. doi:10.1016/j.procs.2017.10.08

14. Ahmed, S. T., Akhand, M. A. (2016). Bangladeshi sign language recognition using fingertip position. 2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec). https://doi.org/10.1109/meditec.2016.7835364