

Real time Object detection in Bangla Sign Language Using Adversarial Deep learning

BY

**Fatema Alim Al Rafika
ID: 191-15-12813**

**Sharthok Saha
ID: 191-15-12700**

AND

**Md. Amran Hossain
ID: 191-15-12751**

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

Supervised By

Ms Shahrin Akter Khushbu
Lecture
Department of CSE
Daffodil International University

Co-Supervised By

Md Ferdouse Ahmed Foysal
Lecturer
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

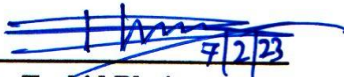
DHAKA, BANGLADESH

FEBRUARY 2023

APPROVAL

This Project titled “**Real time Object detection in Bangla Sign Language Using Adversarial Deep learning.**”, submitted by Fatema Alim Al Rafika, Sharthok Saha and Md Amran Hossain to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 02-02-2023.

BOARD OF EXAMINERS



Dr. Touhid Bhuiyan

Professor and Head

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman



Dr. Sheak Rashed Haider Noori

Professor and Associate Head

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Md. Sazzadur Ahamed

Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Dr. Md. Sazzadur Rahman

Associate Professor

Institute of Information Technology
Jahangirnagar University

External Examiner

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Shahrin Akter Khushbu, Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

For
Jaysal

Shahrin Akter Khushbu

Lecturer

Designation

Department of CSE

Daffodil International University

Co-Supervised by:

Jaysal

Md Ferdouse Ahmed Foysal

Lecturer

Department of CSE

Daffodil International University

Submitted by:

Rafika

Fatema Alim Al Rafika

ID: 191-15-12813

Department of CSE

Daffodil International University

Sharthok

Sharthok Saha

ID: 191-15-12700

Department of CSE

Daffodil International University

Amran

Md Amran Hossain

ID: 191-15-12751

Department of CSE

Daffodil International University

ACKNOWLEDGEMENT

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

We are really grateful and wish our profound indebtedness to **Shahrin Akter Khushbu, Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Deep learning*” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Touhid Bhuiyan**, Professor, and Head, Department of CSE, for his kind help to finish our project and also to other faculty members and the staff of CSE department of Daffodil International University.

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

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

ABSTRACT

Instead of spoken words, sign languages use the visual-manual modality to communicate meaning. Manual articulation and non-manual markers are used to convey meaning in sign languages. It is used by dumb or silent, blind and disabled people all over the world. Having their own grammar and lexicon, sign languages are completely natural languages. Therefore, a machine translator is required to enable them to communicate with the broader public. Computer vision technologies are now well known for helping people translate their languages so that everyone can comprehend them. We have used deep learning methods to detect Bangla sign language. Here we used a custom made data set of 2100 images of our 42 beloved Bangla words. We employed Convolutional Neural Network (CNN) models to classify the words. These models deliver results for photo classification that are more precise. Before these models can be used, the image data must be processed. The employment of particular techniques is required for data preparation. The choices include RGB conversion, filtering, resizing and scaling, and categorization. After using these techniques, image data is preprocessed and made ready for classifier algorithms.

TABLE OF CONTENTS

CONTENTS	Page No
Approval Page	i
Declaration	ii
Acknowledgement	iii
Abstract	iv
List of Tables	v
List of Figures	vi
CHAPTER	
Chapter 1: Introduction	
1.1 Introduction	1
1.2 Motivation	1-2
1.3 Rationale of the Study	2
1.4 Research Questions	2
1.5 Expected Output	3
1.6 Report Layout	3

Chapter 2: Background	Page No
2.1 Related Works	4
2.2 Scope of the Problem	5
2.3 Challenges	5
Chapter 3: Research Methodology	
3.1 Data Collection Procedure	6
3.2 Data Sample	7-8
3.3 Data Processing	9-10
3.4 Proposed Methodology	10-15
3.5 Implementation Requirements	17
Chapter 4: Experimental Results and Discussion	
4.1 Experimental Setup	18
4.2 Experimental Results & Analysis	18-20
4.3 Discussion	20
Chapter 5: Impact on Society, Environment and Sustainability	
5.1 Impact on Society	21
5.2 Impact on Environment	21
5.3 Ethical Aspects	21

Chapter 6: Summary, Conclusion, Recommendation and Implication for Future Research	Page No
6.1 Summary of the Study	22
6.2 Conclusions	22
6.3 Implication for Further Study	22
References	23-24

LIST OF FIGURES

FIGURES	Page No
Figure 1: VGG-16 Model structure	11
Figure 2: Vgg-19 Model structure	12
Figure 3: Resnet-50 Model structure	13
Figure 4: Alexnet Model structure	14
Figure 5: 3D CNN Model structure	15
Figure 6: Workflow diagram	16
Figure 7: Implemented CNN models training and validation performance	19-20

LIST OF TABLES

Tables	Page No
Table 1: Demonstrates the categories, sources of data, and samples.	6
Table 2: After augmentation dataset information	10
Table 3: Model summary of our implemented model	15
Table 4: Model Accuracy comparison	18

CHAPTER 1

Introduction

1.1 INTRODUCTION

Deaf and dumb people suffered with negligence for decades as most general people have no knowledge about sign language. The field of computer vision is expanding to aid people in every conceivable way. Recently, others have started using it. By facilitating the sign language detecting method, to aid the deaf community. There is no international version of sign language. Even though English is the official language of both nations, it is interesting to note that the American and British sign languages (BSL) are not identical. Other sign languages include Indian sign language, French sign language, Japanese sign language, and Bangla sign language (BdSL). The origins of each sign language are distinct and separate. People in contemporary society are working to improve the way of life for the hearing-impaired community by offering alternate forms of communication, education, culture, and sports. These individuals actually have a remarkable amount of talent. A hearing-impaired person and a hearing-normal person or a dumb person and a normal verbal speaker can instantly communicate thanks to automated recognition of sign language (ARSL). A suitable language translator can also be used in conjunction with the ARSL to operate implicitly as an instant interface between those who have difficulty speaking or hearing and non-language-impaired people. This is because sign language has no worldwide form. As a result, experts in the fields of artificial intelligence and natural language processing are working to create a system that can recognize sign language automatically.

1.2 MOTIVATION

According to WHO, around 466 million people have a hearing disability that is over 5% of the world's population. 0.38% of the total population of Bangladesh have speech and hearing disabilities, according to the National Census 2011 [1]. Even though Bangla is the sixth most widely used language in the world. However, the research on BdSL recognition did not progress compared to other sign languages, but recently some researchers are working to make a change [2]. We wanted to use regular words because

as we started learning about sign language, we came to understand that a signed word in Bangla isn't a sequence of signed alphabets. That's when we decided to do deep learning on words rather than alphabets despite the fact that there was no available dataset.

1.3 RATIONALE OF THE STUDY

Acquiring a new language opens the door to a world of new experiences. As students study a new language, they develop an understanding and appreciation of other people, cultures, beliefs, and ways of life, while also developing a deeper understanding of their own culture and personal identity. They learn new ways to think, learn, and communicate with others, and gain a new perspective on their experiences and the world around them. The study of Bangla Sign Language (BDSL) rewards students with these and other benefits. The study of BDSL supports many careers and professions. In medicine, dentistry, the hospitality industry, education, and other career areas, the ability to communicate easily with Deaf adults and children through BDSL is a great asset. It is becoming increasingly important for organizations that provide services to the Deaf community to have employees who are proficient in BDSL. The study of BDSL not only develops the knowledge, skills, and attitudes needed to understand and communicate effectively in BDSL but also expands students' knowledge of language learning in general. In using BDSL to create and convey meaning, students can discover new ways to express their individuality. Communicating in authentic situations in another language increases self-confidence in communication skills, enhances students' critical thinking, and promotes respect for others regardless of differences.

1.4 RESEARCH QUESTIONNAIRES

After we set our heart on working with sign language the questions we came up with was 1.How do spoken languages influence sign languages? 2.What are the latest advancements in computer-based SignLanguage Interpretation? 3.Sign language implementation using real time detection is possible? 4.Existing Dataset of Bangla Sign Language?

1.5 EXPECTED OUTPUT

Our final goal is that we will be able to build a software where it can detect a deaf or dumb person's signed sentences and convert it into text or verbal sound. For now we want to incorporate as many words as possible in our data set and train our models to detect words with precision. In this project, we have only worked on word detection not alphabet detection like most other research, for this our dataset will consist of word sign.

When the image is taken into account, the environment is crucial. A domain can be limited or unconstrained based on illumination conditions because accuracy depends on the environment that affects the data processing step. Data collected in an unfriendly setting introduces impurity that degrades accuracy and is the reason accuracy varies across practice models. For example Md. Uddin et al., proposed a framework using Support Vector Machine (SVM) to recognize BdSL [3].

So, after gathering the data, our top objective will be to properly process it so that it fits the model we prefer. Many useful methods for recognition are used when entering image data. The feature extraction stage follows using a variety of methods.

Our task has been broken down into several sub-tasks. During the data collecting phase, we choose data for train, test, and validation. Where the validation division is used to measure how accurately the model is trained on train data, and the train division is used to train the model. The model performance is then tested using the test division. The data is then processed to suit the model. Apply the artificial neural network model that we have suggested, after all. Our proposed model is built using a Convolutional Neural Network (CNN) model, in addition to several pre-trained models that are also based on.

1.6 REPORT LAYOUT

There are three additional sections in this paper: Backgrounds (which describes the literature review and related works), Research Methodology (which describes all working procedures), Result and Discussion (which discusses and compares all the results of our proposed models), and Conclusion and Future Work (which provides a summary and outlines future work).

Chapter 2

Background

2.1 RELATED WORKS AND LITERATURE REVIEW

This section of our study will provide some analysis of documents where various strategies have been used to produce the desired results.

Oishee Binte Hoque [4] proposed faster R-CNN. Their training data is put into the convolutional neural network, which generates a feature map depending on the model's aspect ratio and sizes as previously discussed in the Faster R-CNN overview. The feature map then sends regions to the RPN that have a chance of becoming a sign gesture or a letter. In order to categorize each proposal into a defined number of classes, ROI pooling resizes the feature map into fixed sizes for each proposal. Md Shafiqul Islam [2] For the purpose of recognizing Bangla sign language, they compared the effectiveness of the suggested CNN model with one of the company's most widely used models, LeNet-5. The Bangla numerals and Bangla sign language alphabet had accuracy rates of 98.73% and 99%, respectively, thanks to the adoption of the LeNet-5 architecture. LeNet5 training takes just 11 minutes and 6 minutes on the architecture, compared to the suggested model's 18 minutes and 9 minutes, respectively, for alphabet and numeric training. As a consequence, the proposed model outperformed the corresponding LeNet-5 model in terms of results. However, compared to the LeNet-5 architecture, which is quite simple, this training takes a few minutes longer.

Sherin Sultana Shanta [5] used SIFT with CNN and saw a massive increase in accuracy. SIFT helps CNN to extract features without scaling, resolution and rotation problems.

Dipon Talukder [1] used YOLOV4 and got 97.56% accuracy in detecting alphabets. They also tried to use punctuation but the problem is they tried to create words by syncing alphabets which is not a usual way of talking among the disabled people. Every word has its own sign rather than a sum of alphabet signs.

Through data augmentation, 15% of the Ishara-Lipi sign character database's data are preserved for testing and 85% are kept for training. After 50 iterations, the model for the Ishara-Lipi sign character database achieves 92.65% accuracy on the training set and 94.74% accuracy on the validation set [7]. The authors used ADAM optimizer with a

learning rate of 0.001. The model has a 9-layer CNN. For convolution 1 and 2, where filter size is 32, kernel size is (5x5), Stride is (1x1), same padding with ReLU (1) activation. Followed by a 5 x 5 max-pooling layer. Then used 25% dropout to reduce overfitting. This is just an open dataset of Bangla alphabet and digit signs. Bikash Chandra Karmokar [8] used Neural Network Ensemble and NCL. They got 93% with NCL and 10 NNs with feature extractions. The authors suggested the experimental setting shouldn't have any hues that clash with the color of human skin. A higher pixel count on the camera is necessary for improved image quality. When taking the pictures, there should be enough light.

Rahat Yasir [9] has related most of the problems we faced but again only with alphabets. Their segmentation of RAW Image dataset and test result (Original pics-hidden Nodes-50) has very poor accuracy which we faced. On the other hand when they used gray images instead of RGB images with PCA features-hidden Nodes-50, the segmentation of PCA dataset and test result came with great accuracy.

2.2 SCOPE OF THE PROBLEM

Use Dynamic Loading for Dataset: Our original dataset was quite large and is impossible to use without a server with a lot of RAM and disk space. A possible solution is to split the file names into training, validation, and test sets and dynamically loading images in the Dataset class. Using such a loading technique would allow us to train the model on more samples in the dataset.

2.3 CHALLENGES

The challenge we faced most was collecting the images because there is no existing dataset for Bangla words. Also the raw created a problem for us while building the models. Also lack of GPU was a big problem for us.

Chapter 3

Research Methodology

This section will go over the instrumentation that was utilized, how the data were collected, how they were analyzed, and what model was suggested. For our suggested work, we employ the deep learning classification technique. For this categorization, we used a multi-layer Convolutional Neural Network (CNN) [5-7].

3.1. DATA COLLECTION

For the purposes of the suggested study, we amassed a total of 3280 data by ourselves while observing a Bangla Sign Language instructor. We gathered unprocessed data from the authors' families and other well-known individuals for each of the 41 classifications. We have taken data from different age group, gender and skin color to make our dataset rich and more realistic. Though skin color became a problem for us later to match with backgrounds.

Table 1 demonstrates the categories, sources of data, and samples.

Number of class	Number of raw images per class	Source of images	Total Images
41	80	Custom	3280

With the soaring demand for computing power and storage, it is challenging to deploy deep neural network applications. Consequently, while implementing the neural network model for computer vision, a lot of effort and work is put in to increase its precision and decrease the complexity of the model. CNNs represent many challenges such as overfitting, exploding gradients, class imbalance and the need of large datasets.

Sample

দুপুর (Afternoon)



আপনার (Your)



বাবা (Father)



খারাপ (Bad)



সুন্দর (Beautiful)



বোন (Sister)



ছেলে (Boy)



বৌদ্ধ (Buddhist)



পছন্দ (Choice)



শীত/ঠান্ডা (Cold)



সর্দি-কাশি (Cough)



ডাল (Pulses)



ডাক্তার (Doctor)



ভয় (Fear)



জ্বর (Fever)



মাছ (Fish)



ফুফু (Aunty)



মেয়ে (Girl)



ভালো (Good)



হিন্দু (Hindu)



হাসপাতাল (Hospital)



গ্রীষ্মকাল (Summer)



বাড়ি (House)



স্বামী (Husband)



ইসলাম (Islam)



মা (Mother)



মামা (Uncle)



মাংস (Meat)



নামাজ (Namaz)



রাত (Night)



ব্যথা (Pain)



পুলিশ (Polish)



বৃষ্টি (Rain)



রোজা (Roja)



লবণ (Salt)



অসুস্থ (Sick)



ঘুম (Sleep)



চা (Tea)



তোমার (Your)



ভাই (Brother)



শাক-সবজি
(Vegetable)



3.2. DATA PREPROCESSING

Following data collection, we divide the data up into train, validation, and test folders, as shown in table 2. Although sign language places a lot of emphasis on expressions, we primarily want our data to learn from hand gestures. For this reason, we also added additional facial features like a nose, lips, an eyebrow, etc. Because the model will only learn the features it requires when the data is fed into it. The overfitting issue will then arise when our model picks up a lot of new features unrelated to our research.

Image resize: All photos must be scaled to a fixed size before being input to the CNN since neural networks require inputs of the same size. Less shrinking is needed the larger the fixed size. Less reduction results in less distortion of the image's internal characteristics and patterns.

Pixel brightness transformations: Position-dependent brightness correction and grayscale transformation are the two different kinds of pixel brightness transformations. Regardless of the position of the pixel, grayscale transformation changes the brightness of the pixel. Transformations to grayscale are typically utilized when a human will be able to view the results.

Dividing into word classes: We have divided our images in 41 classes according to the signs.

Labeling: When you annotate particular items or features in an image, this is known as image labeling. Computer vision models learn from picture labels how to recognize a specific object in an image.

Augmentation: Data augmentation uses existing data to create modified copies of datasets, which are then used to artificially increase the training set. It involves making little adjustments to the dataset or creating new data points using deep learning. As we first didn't get our desired accuracy, we had to augment our data to make it work.

Table 2: After augmentation dataset information

Number of class	Augmented image per class	Image per class after augmentation	Total Images in dataset
41	100	180	9430

3.4. IMPLEMENTED CNN MODEL

A CNN [10] is comprised of convolutional layers followed by one or more fully connected layers as in a standard multilayer shallow neural network . In CNN, each convolutional layer learns features from the input data making this architecture well suited to process images. CNNs manipulate multiple hidden layers for learning to capture different features from input data. The complexity of the learned data features may increase for every hidden layer. The four CNN model we proposing are-

VGG 16: In their publication "Very Deep Convolutional Networks for Large-Scale Image Recognition," K. Simonyan and A. Zisserman from the University of Oxford introduced the convolutional neural network model known as VGG16. In the top five tests, the model performs 92.7% accurately in ImageNet, a dataset of over 14 million images divided into 1000 classes. It was a well-known model that was submitted to ILSVRC-2014. By sequentially substituting several 3x3 kernel-sized filters for AlexNet's big kernel-sized filters (11 and 5, respectively, in the first and second convolutional layers), it improves upon AlexNet. Using NVIDIA Titan Black GPUs, VGG16 underwent weeks of training. Each filter has the same padding and has 512 filters of size (3, 3). The stack of two convolution layers then receives this image. As opposed to AlexNet's and ZF-11*11 Net's and 7*7 filters, we use 3*3 filters in these convolution and max-pooling layers. It also employs 1*1 pixels in some of the layers to adjust the amount of input channels. To prevent the spatial characteristic of the image, 1-pixel padding is applied after each convolution layer.

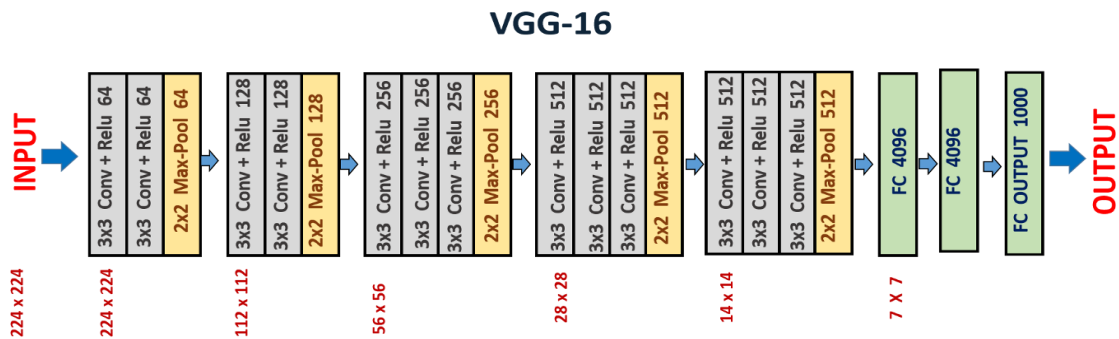


Figure 2: VGG-16 Model structure

The input to cov1 layer is of fixed size 224 x 224 RGB image. Convolutional (conv.) filters were applied with a very small receptive field, 33 (which is the smallest size to capture the notion of left/right, up/down, and center). The picture is then passed through the stack of convolutional (conv.) layers. It also uses 11 convolution filters in one of the setups, which may be thought of as a linear transformation of the input channels (followed by non-linearity).

VGG-19: A variation of the VGG model called VGG19 has 19 layers in total (16 convolution layers, 3 Fully connected layer, 5 MaxPool layers and 1 SoftMax layer). This network received a fixed-size (224 * 224) RGB picture as input, indicating that the matrix was shaped (224,224,3). They were able to cover the entirety of the image by using kernels that were (3 * 3) in size with a stride size of 1 pixel.

The VGG19 model (also known as VGGNet-19) has the same basic idea as the VGG16 model, with the exception that it supports 19 layers. The numbers "16" and "19" refer to the model's weight layers (convolutional layers). In comparison to VGG16, VGG19 contains three extra convolutional layers.

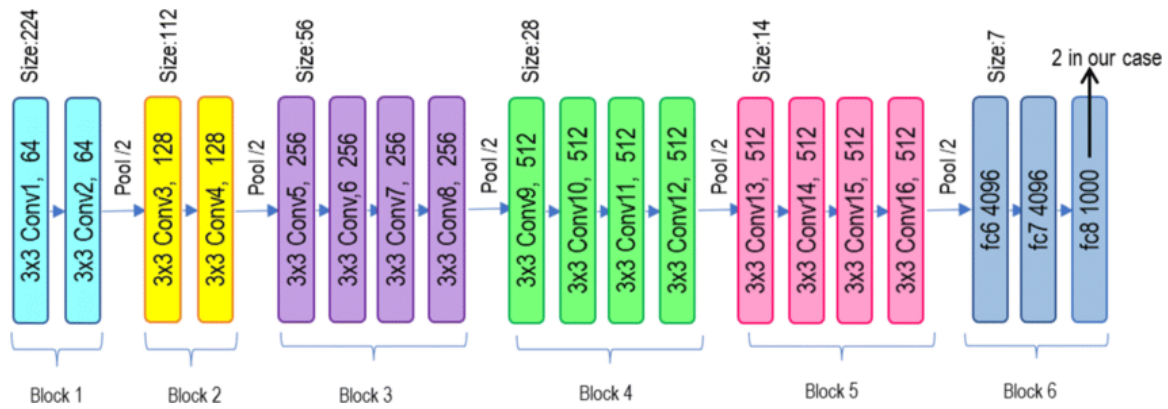


Figure 3: Vgg-19 Model structure

A fixed size of (224 * 224) RGB image was given as input to this network which means that the matrix was of shape (224,224,3). The only preprocessing that was done is that they subtracted the mean RGB value from each pixel, computed over the whole training set. Used kernels of (3 * 3) size with a stride size of 1 pixel, this enabled them to cover the whole notion of the image. spatial padding was used to preserve the spatial resolution of the image.

RESNET: The Residual Blocks idea was created by this design to address the issue of the vanishing/exploding gradient. We apply a method known as skip connections in this network. The skip connection bypasses some levels in between to link layer activations to subsequent layers.

This creates a leftover block. By stacking these leftover blocks together, resnets are created. The strategy behind this network is to let the network fit the residual mapping rather than have layers learn the underlying mapping. Thus, let the network fit instead of using, say, the initial mapping of $H(x)$,

$$F(x) := H(x) - x \text{ which gives } H(x) := F(x) + x.$$

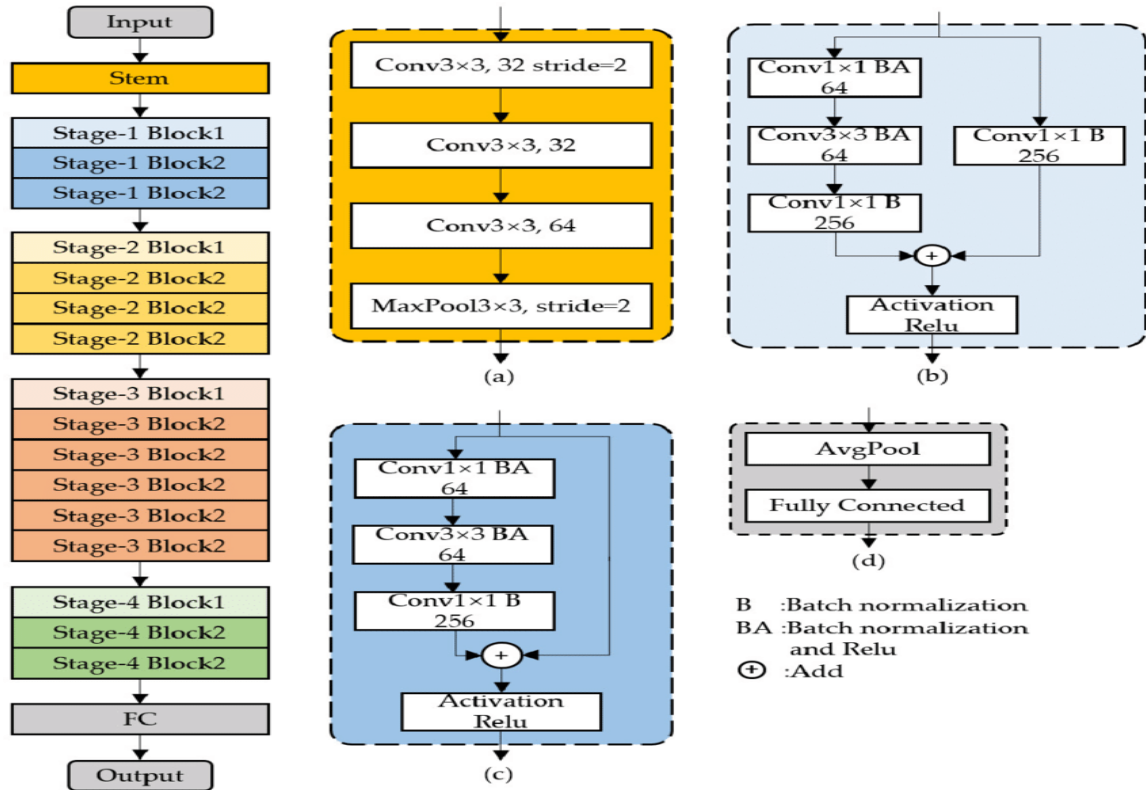


Figure 4: Resnet-50 Model structure

Alexnet: After competing in ImageNet Large Scale Visual Recognition Challenge, AlexNet shot to fame. It achieved a top-5 error of 15.3%. This was 10.8% lower than that of the runner up. The primary result of the original paper was that the depth of the model was absolutely required for its high performance. This was quite expensive computationally but was made feasible due to GPUs or Graphical Processing Units, during training. There are eight learnable layers in the Alexnet. Relu activation is used in each of the five levels of the model, with the exception of the output layer, which uses max pooling followed by three fully connected layers. One thing to keep in mind is that the authors used padding because Alexnet has a deep architecture to prevent the size of the feature maps from dwindling significantly. The photos with a size of 227X227X3 are used as the model's input. The input to the Model is RGB images.

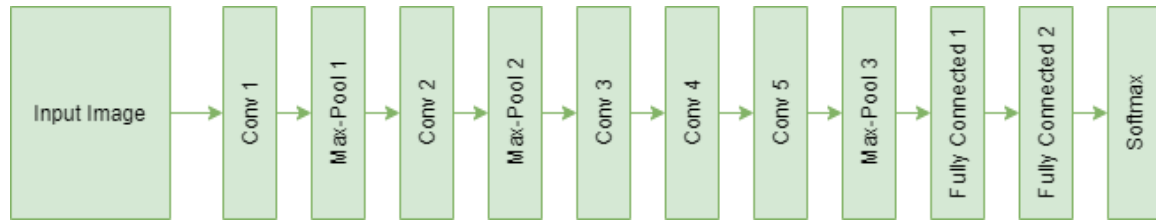


Figure 5: Alexnet Model structure

3D CNN: As a time-consuming procedure that currently necessitates expert analysis, 3D CNNs are being developed to enhance the identification of moving and 3D pictures, such as video from security cameras and medical scans of malignant tissue.

Due to their intricacy, 3D CNN development is still in its infancy, but the advantages they can provide are ones you should familiarize yourself with. Continue reading to discover how this new area raises the bar for deep learning in computer vision.

Convulsion layer is Kernels, a filter layer made up of learnt parameters, convert images into processable data in this layer. Each study employs many kernels, each of which filters for a distinct property.

Pooling layer, The convolution-generated activation maps are pooled, or downsampled. Similar to the convolution process, pooling involves moving a filter across an activation map and analyzing a tiny section at a time.

Fully Connected (FC) Layer is The output layers are flattened, the probabilities found are examined, and the result is given a value, a logit, after several iterations—sometimes thousands—of convolution and pooling. The Fully Connected layer does this analysis, processing each flattened output layer through interconnected nodes in a manner akin to a fully connected neural network (FCNN).

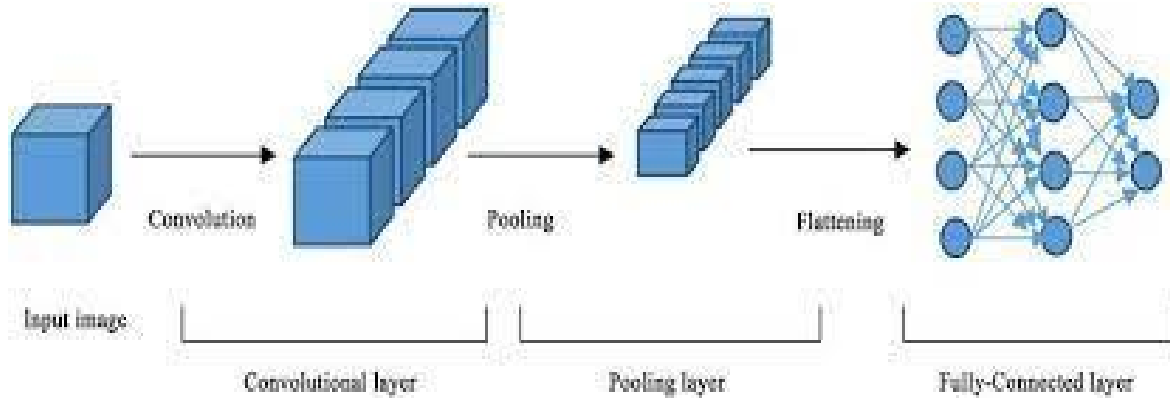
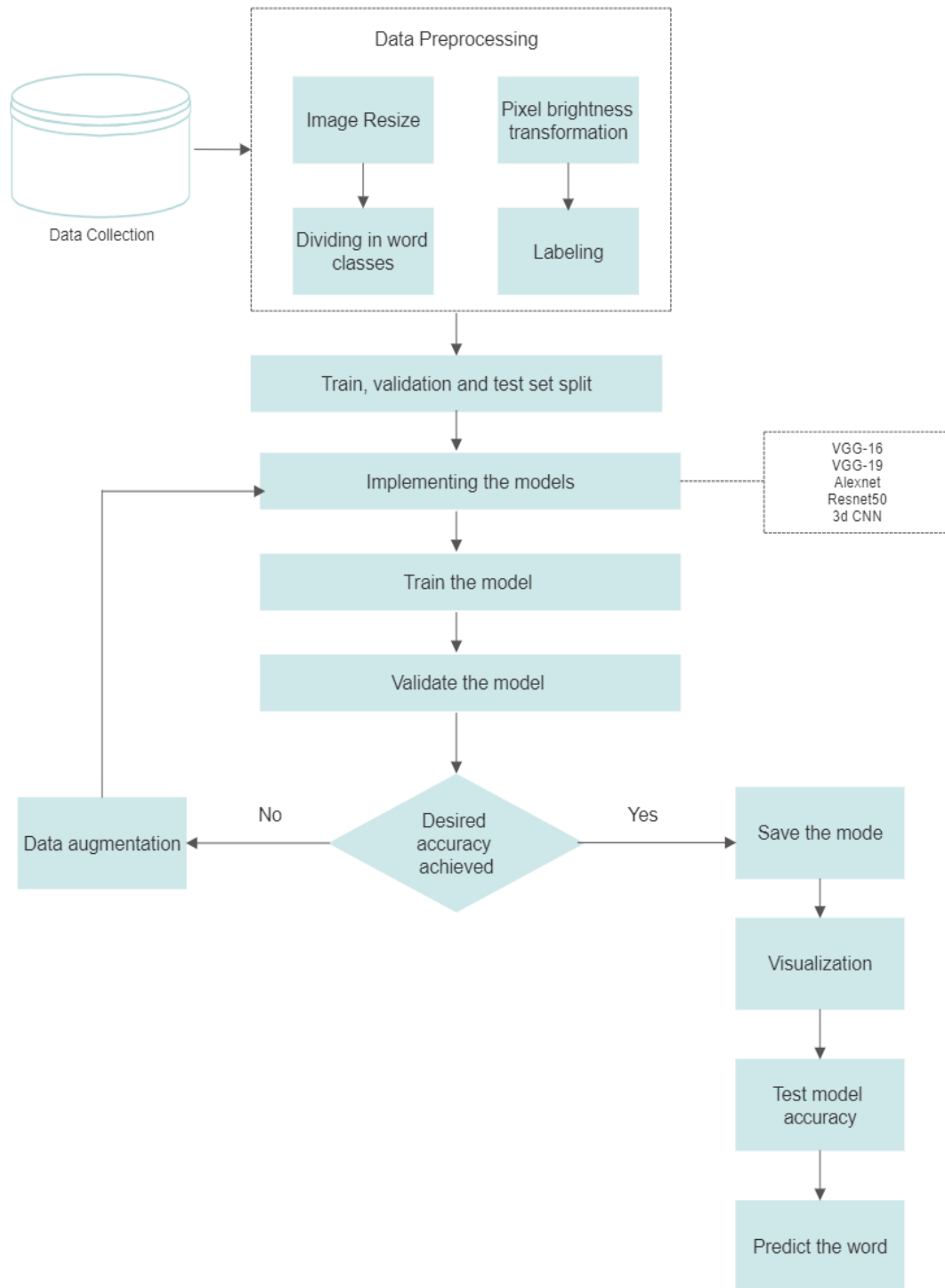


Figure 5: 3D CNN Model structure

Table 3: Model summary of our implemented model

Model Name	Total params	Trainable params	Non-trainable Params
VGG-16	15,743,337	1,028,649	14,714,688
VGG-19	21,053,033	1,028,649	20,024,384
Alexnet	28,120,793	28,099,657	21,136
Resnet-50	27,802,538	4,214,826	23,587,712
3d CNN	264,297	264,105	192

Figure 6: Workflow diagram



3.5 IMPLEMENTATION REQUIREMENTS

To implement this project we needed Web IDE, NVIDIA high GPU, high RAM. First we faced a lot of problems with the initial IDE we chose that is Google Colab, as we faced a lot of runtime expiry. Then we opted for Google Colab pro to use High RAM runtime. Still we face GPU runout problems.

To run the models we ran followed below steps:

Import the Libraries, Initialize our model, Convolution Step, Pooling Step, Add a second convolution layer and pooling layer, Flatten Step, Full Connection layer, Compiling the CNN.

CHAPTER 4

Experimental Results and Discussion

4.1 EXPERIMENTAL SETUP

For this experiment we have collected Bangla signs for 41 words and put them through various models to see which models act the best according to our expectation. We trained our model using different epochs to make it work better.

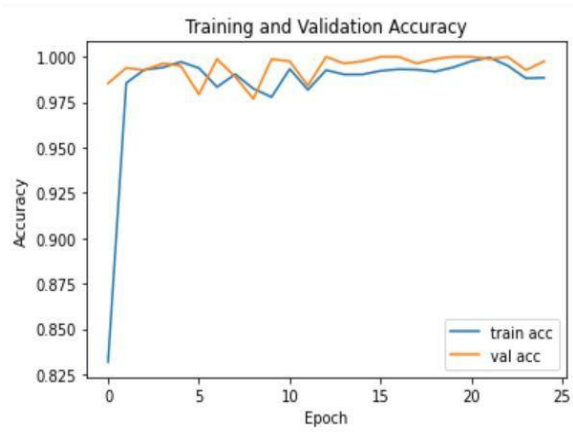
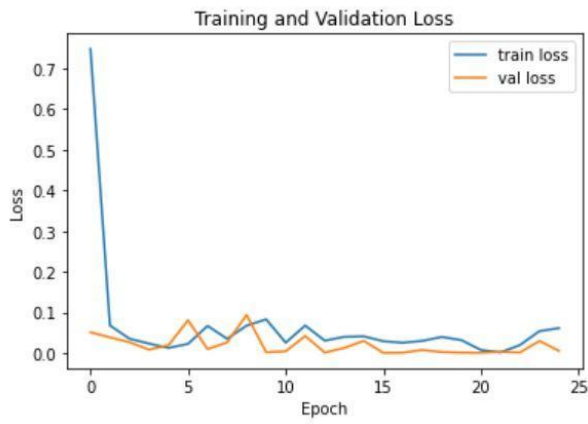
4.2 EXPERIMENTAL RESULTS & ANALYSIS

The findings of our experiment were quite satisfactory to us. As it shows with proper dataset and training models can detect many words. Though the dataset still needs some polishing but right now the results are quite good below is our outcome of different models. We have implemented VGG-16, VGG-19, Alexnet-50, Resnet-50, 3D CNN. The best accuracy we got is in VGG-16 that is 87%. The rest of our models worked pretty nicely as well. Now we can proceed to the next level.

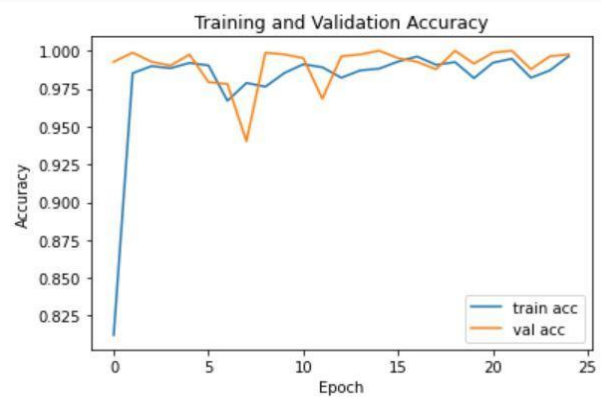
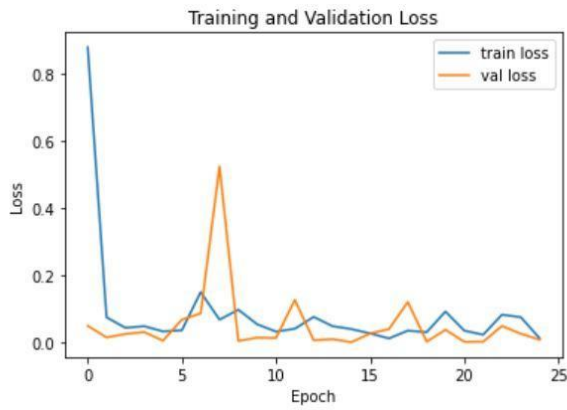
Table 4: Model Accuracy comparison

Model Name	Accuracy
VGG-16	87%
VGG-19	86%
Alexnet	81%
Resnet50	79%
3d CNN	91%

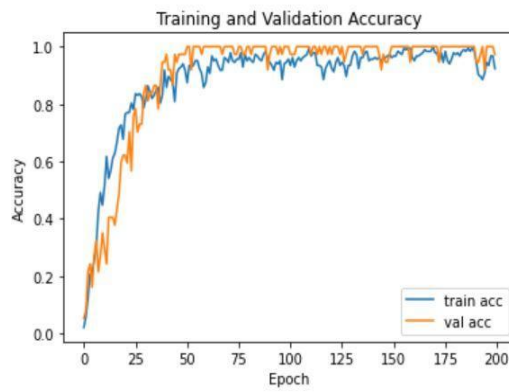
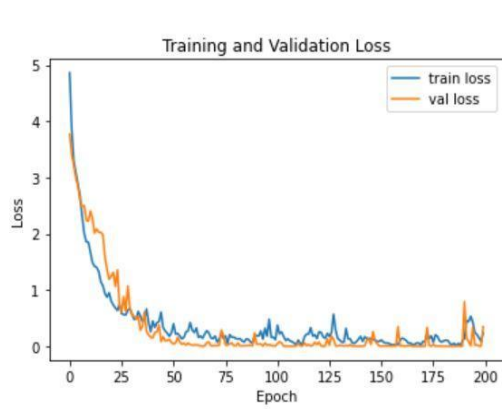
Figure 7: Implemented CNN models training and validation performance



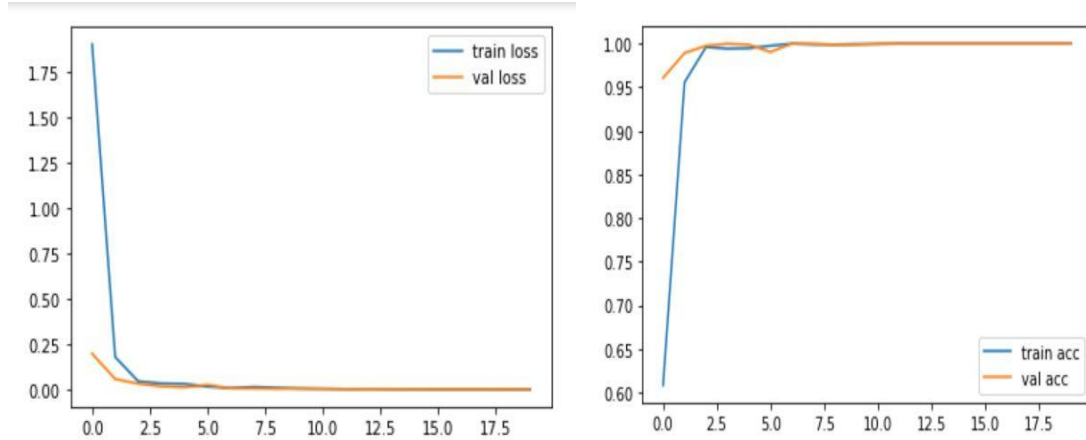
VGG-16



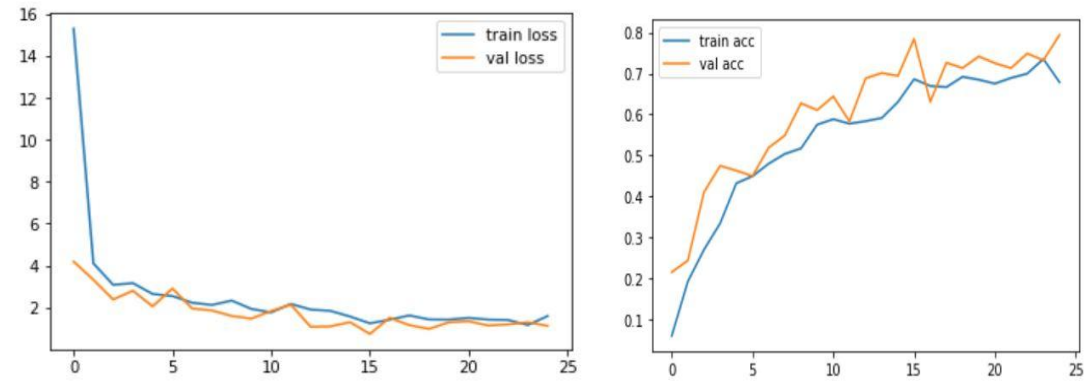
VGG-19



3D CNN



Alexnet



Resnet50

4.3 DISCUSSION

A very minor portion of sign language detection is covered by this work. To reach the ultimate objective of closing the communication gap between the hearing impaired and the remainder, much more work needs to be done and can be done in this area.

To increase the accuracy of the recognition of lone indicators, we must improve the work we did for our thesis. In order to distinguish between signs with identical hand movements and positions utilizing the already collected information, features linked to hand shapes must first be extracted. The following stage will be to increase the quantity of training samples and experiment with training techniques other than neural networks in order to determine which one yields the best results.

Chapter 5

Impact on Society, Environment and Sustainability

5.1 IMPACT ON SOCIETY

Our project will only have a good impact on society as this project doesn't hurt anyone's belief, religion, caste, personality, community, nationality. This will make the dumb and deaf people feel more incorporated into the society they have lived in for years neglected, deprived. This will make their very harsh life a tiny bit easier.

5.2 IMPACT ON ENVIRONMENT

This will have no bad impact on the environment as this experiment doesn't need to use any chemical or radiation of any sort. This experiment is totally cruelty free and doesn't harm any human or animal in the process. This experiment will require no connection to nature let alone harming it.

5.3 ETHICAL ASPECTS

The ethical issues we might face:

- This technology may be seen as assimilation and caving in to the expectations of the privileged community, but it actually helps the disabled community integrate even more fully into the abled world. This might result in less work.
- The dataset must be sufficiently diverse to account for people of all skin tones and in all environments, in order to accommodate deaf people. A bias in the data may harm deaf people of a particular ethnic group.

Chapter 6

Summary, Conclusion, Recommendation and Implication for Future Research

6.1 SUMMARY OF THE STUDY

If we summarize our research we can say that sign language is a very important part of hearing disabled people's life. The more we work on Bangla sign language the more we can show the world the sweetness of it. We need to incorporate more models, specially if possible hybrid deep learning models to make the words detected better. Our government can also take a note and make opportunities for the data scientist to work in this field.

6.2 CONCLUSIONS

Using methods from deep learning, we implemented a system for real-time sign language gesture identification in this research. We discovered that occasionally straightforward solutions outperform more complex ones. The rather simple skin segmentation model ended up being the best skin masks despite attempts to employ a clever segmentation algorithm. We also understood the challenges and limitations of building a dataset from the start. Looking back, it would have been convenient to have a dataset to work with already. For contradicting reasons, some letters were more difficult to categorize in our real time testing.

6.3 IMPLICATION FOR FURTHER STUDY

Due to the size of our initial dataset, using it requires a server with plenty of RAM and disk space. The division of the file names into training, validation, and test sets and the dynamic loading of photos in the dataset class are potential remedies. We could train the model on more samples in the dataset if we used such a loading strategy.

References

- [1] D. Talukder and F. Jahara, "Real-Time Bangla Sign Language Detection with Sentence and Speech Generation," *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, 2020, pp. 1-6, doi: 10.1109/ICCIT51783.2020.9392693.
- [2] M. S. Islalm, M. M. Rahman, M. H. Rahman, M. Arifuzzaman, R. Sassi and M. Aktaruzzaman, "Recognition Bangla Sign Language using Convolutional Neural Network," *2019 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, 2019, pp. 1-6, doi: 10.1109/3ICT.2019.8910301.
- [3] M. A. Uddin and S. A. Chowdhury, "Hand sign language recognition for Bangla alphabet using Support Vector Machine," *2016 International Conference on Innovations in Science, Engineering and Technology (ICISSET)*, 2016, pp. 1-4, doi: 10.1109/ICISSET.2016.7856479.
- [4] P. P. Urmee, M. A. A. Mashud, J. Akter, A. S. M. M. Jameel and S. Islam, "Real-time Bangla Sign Language Detection using Xception Model with Augmented Dataset," *2019 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)*, 2019, pp. 1-5, doi: 10.1109/WIECON-ECE48653.2019.9019934.
- [5] Albawi, Saad, Tareq Abed Mohammed, and Saad Al-Zawi. "Understanding of a convolutional neural network." In *2017 International Conference on Engineering and Technology (ICET)*, pp. 1- 6. Ieee, 2017.
- [6] Rahaman, Muhammad Aminur et al. "Bangla language modeling algorithm for automatic recognition of hand-sign-spelled Bangla sign language." *Frontiers of Computer Science* 14 (2019): n. pag.
- [7] M. Sanzidul Islam, S. Sultana Sharmin Mousumi, N. A. Jessan, A. Shahariar Azad Rabby and S. Akhter Hossain, "Ishara-Lipi: The First Complete Multipurpose Open Access Dataset of Isolated Characters for Bangla Sign Language," *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, 2018, pp. 1-4, doi: 10.1109/ICBSLP.2018.8554466.
- [8] Chandra Karmokar, Bikash & Alam, Kazi & Siddiquee, Md. (2012). Bangladeshi Sign Language Recognition Employing Neural Network Ensemble. *International Journal of Computer Applications*. 58. 10.5120/9370-3846.
- [9] R. Yasir and R. A. Khan, "Two-handed hand gesture recognition for Bangla sign language using LDA and ANN," *The 8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA 2014)*, 2014, pp. 1-5, doi: 10.1109/SKIMA.2014.7083527.

- [10] R. Vaillant, C. Monrocq and Y. Le Cun, "An original approach for the localization of objects in images," *1993 Third International Conference on Artificial Neural Networks*, 1993, pp. 26-30.
- [13] F. Yasir, P. W. C. Prasad, A. Alsadoon and A. Elchouemi, "SIFT based approach on Bangla sign language recognition," *2015 IEEE 8th International Workshop on Computational Intelligence and Applications (IWCIA)*, 2015, pp. 35-39.

ORIGINALITY REPORT

14%

SIMILARITY INDEX

11%

INTERNET SOURCES

4%

PUBLICATIONS

14%

STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Bournemouth University Student Paper	3%
2	Submitted to Harrisburg University of Science and Technology Student Paper	2%
3	Submitted to An-Najah National University Student Paper	1%
4	Submitted to National College of Ireland Student Paper	1%
5	Submitted to Bangladesh University of Professionals Student Paper	1%
6	Submitted to Liverpool John Moores University Student Paper	1%
7	Submitted to University of Greenwich Student Paper	1%
8	Submitted to The Robert Gordon University Student Paper	1%

9	Submitted to University of Sydney Student Paper	1%
10	Submitted to Camarines Sur Polytechnic Colleges Student Paper	1%
11	Submitted to Higher Education Commission Pakistan Student Paper	<1%
12	Submitted to University of Wales Institute, Cardiff Student Paper	<1%
13	Submitted to College of Science and Technology, Bhutan Student Paper	<1%
14	Timothy Reagan. "Linguistic Legitimacy and Social Justice", Springer Science and Business Media LLC, 2019 Publication	<1%