A Smart Avatar Tutor for Mimicking Characters of Bangla Sign Language

by

Apu Islam
23241069
Somaya Al Sadia Rahman
20101560
Pratick Bhowmick
20101262

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
Brac University
January 2024

Declaration

It is hereby declared that

1. The thesis submitted is my/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:

Apu Islam
23241069

Somaya Al Sadia Rahman
20101560

Pratick Bhowmick
20101262

Approval

The thesis/project titled "A Smart Avatar Tutor for Mimicking Characters of Bangla Sign Language" submitted by

1. Apu Islam (23241069)
2. Somaya Al Sadia Rahman (20101560)
3. Pratick Bhowmick (20101262)

Of Summer, 2015 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on August 23, 2015.

Examining Committee:

Supervisor:
(Member)

Md. Golam Rabiul Alam, PhD
Professor
Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam, PhD
Professor
Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Computer Science and Engineering
BRAC University

Abstract

This paper presents an efficient novel approach to developing a smart avatar tutor for teaching Bangla Sign Language (BSL). The avatar is designed to mimic the numbers and alphabets used in BSL, providing an interactive and engaging tool for learners. The system employs image processing and machine learning techniques to identify and analyze hand gestures and signs in a dataset, generating corresponding animated sequences that the avatar can use to replicate them. Following preliminary pre-processing of the dataset, we collected hand gestures and signs to train the system. The avatar's user interface (UI) is designed to be intuitive and user-friendly, allowing learners to practice physically to engage their sensorimotor skills, making the learning more efficient, and receiving instant feedback on their performance using cosine similarity. The results of a user study, conducted with 10 participants from the National Federation of the Deaf, showed that the avatar was effective in improving learners' BSL skills and was perceived positively in terms of usability and engagement. Though there is a little research on American Sign Language avatar tutors or a learning system for American Sign Language, there isn't any research conducted on a smart avatar tutor for teaching Bangla Sign Language (BSL). The proposed system has the potential to make BSL learning more accessible and enjoyable for a wider audience.

Keywords: Bangla Sign Language; Image processing; YOLOv8; Avatar; Cosine similarity

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

One of the biggest components of human life is communication, it acts as a bridge for human communication and social interactions among people all around the world. Although communication plays a huge role in our social interaction and quality of life as a human being, there are different ways of language through interaction. One may talk verbally, others may use signs or written forms of language as a communication medium. A significant percentage of people in Bangladesh use sign language as their main medium of communication. Bangladesh has 133,517 people with hearing impairments and 194,394 people with voice problems, according to the Department of Social Services [13]. It may seem to be a mere medium of communication to us but for many, it is like a bridge between them and the outside world. These communication techniques or how efficient it will be to connect to the outside world solely depend on how proficient in sign language the hearing-impaired people and others accompanying them are.

To address this special need, our research introduces an effective and innovative technological idea: the ”Smart Avatar Tutor for Mimicking Characters of Bangla Sign Language.” An intelligent avatar that can mimic Bangla sign language can help to teach these people who are without knowledge of sign language. This will set a significant advancement in assistive online educational technology by incorporating the YOLOv8 Nano baseline model. YOLOv8 Nano baseline model is a simpler version of the ”You Only Look Once” series and is vastly known for its effectiveness and precision in object identification and also the lightness of the model, especially in settings with confined processing capacity. This model is very suitable for identifying and analyzing the complex motions of Bangla Sign Language (BSL) due to its real-time object identification capabilities. Given the YOLOv8 Nano baseline model, the avatar is capable of mimicking these actions and will provide users with a more effective, accurate, and entertaining learning tool.

The distinguishing feature of our paper is the unique application of the YOLOv8 Nano Baseline model in the industry of sign language teaching technologies. although there have been other attempts to add technology in sign language teaching platforms mainly using CNNs, RNNs, 3D CNNs, Pose Estimation Models, Combining different types of neural networks, such as CNNs for feature extraction and RNNs/LSTMs for sequence modeling, and so on. However, it showed the highest accuracy using the Yolo model series. Yolov8 performs efficiently in the case of

real-time image detection for sign language. But we choose to use the lightweight model yolov8 nano baseline model for this new learning platform by considering the computational complexity, and storage complexity as a concern.
Furthermore, this paper will reflect on the implications of technological advancement for the Bangla sign language teaching platform, which makes it more accessible to people with any kind of smart device. It will explore how learning Bangla sign language for a person with hearing impairments and mute people will be revolutionized by smart teaching avatars or 2d hand gestures. Along with giving a thorough backdrop for this study, the article will discuss crucial earlier research and detail the precise goals and development processes used in the creation of the smart avatar teacher, with an emphasis on the use of the YOLOv8 Nano baseline model. Though yolov8 provides more accuracy and robustness in object detection by using the complex neural network architecture, due to its large size it is suitable for high-power servers or devices with high-end GPUs. This could be a problem for the users. By keeping that in mind we went for the YOLOv8 Nano Baseline model which is optimized for speed and efficiency as it is best fitted where real-time processing is required on devices with limited computational power, such as mobile apps, home personal computers with less CPU capacity, IoT devices, or any system where minimizing computational load and power consumption is important. This whole system mimics the provided hand gestures for BSL in a 2D model and creates an efficient, fast-performing real-time hand gesture recognition platform that facilitates the user with more exciting and interactive learning with the help of a smart avatar.
This development of smart avatars teaching Bangla sign language can be a breakthrough toward inclusive education especially based on the YOLOv8 Nano baseline model. This is a major step toward people who need special care and facilities in education. This technological revolution can be a major step in shaping our special education system for those people. Through this research, we hope to make a small contribution to the field of adaptive educational technology and encourage further advancement in this significant area.

1.2 Research Area

The study of our paper, "A Smart Avatar Tutor for Mimicking Characters of Bangla Sign Language," converges assistive educational system, public health, and machine learning. it intersects the critical areas.
Based on the recent statistics on global hearing loss, in 2019 approximately 1.757 billion people globally had hearing loss, which is basically in every five people. Among them, 403.73 million people had hearing loss that was moderate and higher in severity after using hearing aids. By 2050, they predicted that a projected 2.745 billion people will have hearing loss, which is a 56.71% increase from 2019. The study provided an overview of the global perseverance of hearing loss and its projected increase.[6] According to another research the world has almost 5% population that either has a hearing or speaking disability, which results in 430 million people worldwide. This number is projected to grow to 900 million in the next 25 to 30 years. The WHO statistics show that 32 million children are acoustically impaired.[8] From those numbers above we can say that it is high time to consider those disabilities as public health concerns. Therefore, our work will improve the quality of hearing and speaking disabled people of our country which will result in improving their quality

of life as well as social integration.

In our paper, we have incorporated advanced gesture recognition models like the yolov8 nano baseline model and a 2D animated hand that will showcase the signs to the learners. It not only helps learners with hearing and speaking disability but also pushes the boundary of assistive educational technology. The avatar provides a visual method of learning that is easy to mimic and learn Bangla sign language at an individual's own pace. Unlike the traditional method, an avatar can be more interactive, engaging, and more patient, so people who feel nervous or self-conscious can easily take lessons online in their comfort zone. Its consistency makes it even better than humans in this case. The avatar will be available at any moment and users can practice as many times as they need and rely on the accuracy of their mimicking as the proficiency of learning sign language.

The base of our research is using a real-time sign language recognition system using advanced and accurate machine learning models. the sign language recognition techniques involve complex computer vision methods and deep learning algorithms which are specially trained on accurate and robust sign language datasets. Implementing real-time processing in both optimized and accurate ways is a challenging part of machine learning and computer vision.to make our system more optimized and easy to use on low-power consumption devices we used the YOLOV8 nano baseline model. Because of its fast-performing ability, it works best in the case of real-time evaluation as a smart avatar.

This paper also contributes to the study of linguistics mainly in understanding and interpreting Bangla sign language. The creation of an online learning platform using smart avatars will contribute to the preservation and standardization of the Bangla language. We can further improve our work by including different learning levels such as alphabets, words, sentences, etc. It can be considered as the documentation of the Bangla sign language for future accessibility of this language. The access to Bangla sign language is very limited. This could be an opportunity to make it more accessible. This intersects with computational linguistics by incorporating technology with language, a rarely explored area in computational linguistics.

The development of an avatar tutor needs a user-oriented design that will focus on the specific needs and preferences of hearing impairments and mute people. This will require a good understanding of human-computer interaction. It perfectly aligns with HCI research focused on more natural ways for learners to interact with technology. As it provides real-time feedback besides teaching it enhances its interactive experience. The system combines visual and textual elements side by side and this multimodal interaction is an important part of HCI research. HCI research also explores how virtual animated objects can impact user experience, motivation, and learning outcomes. This study further will provide valuable insights into creating more engaging and humane systems for users with special educational and communication needs.

An integral part of this research is the evaluation of the accuracy of the user's hand gestures mimicking the 2D avatar. By tracking the proficiency from the accuracy data the system can later on assess how effective the system is in terms of interacting and teaching people sign language. From the performance curve of this, we can make further modifications or address where the problem lies. The system can track metrics of user engagement, motivation, experience, frequency of use, session duration, or progress of the learning. This can further show us the effectiveness of educational technologies and make them more engaging for learners if needed.

As a whole, our study covers multiple fields and makes significant advancements in the fields of assistive educational technology, machine learning, computer vision, linguistics and sign language interpretation, human-computer interaction, public health, education technology evaluation, and so on. A smart avatar tutor for Bangla Sign Language using the YOLOv8 Nano model can solve the communication and social interaction gap in the rising hearing impairments and mute society in the upcoming years.

1.3 Background Information

Sign language has been a mode of communication among the deaf community for decades. As modern technology evolved, to make the life of those people easier, recognition of sign language emerged. Visual-gestural languages with names like sign languages are distinguished by their currently held grammar, lexicon, and syntax. The way vocal language plays an important role in expressing our feelings with other people, sign language also plays a crucial role in facilitating communication, expression, and social integration for individuals with hearing impairments. In Bangladesh, over 200,000 people use Bangla Sign Language to communicate.[11] However, the rest of the people in the country hardly have any idea regarding communication through the Bangla sign language. So, there is a situation where those people who are bound to communicate through sign language struggle to express their words which is very frustrating. Furthermore, those who also want to understand the Bangla Sign Language don't have many opportunities to learn it because most of the previous studies aren't that effective in terms of teaching sign language more efficiently and simply. So, if something can help to teach sign language, they will be able to communicate with the people who use Bangla Sign Language to express their words. This will be able to connect and act as a bridge between the people with hearing impairment and the rest of the people in Bangladesh. The paper we are working on focuses on developing a system where an avatar teaches bangla sign language to make communication easier for those people with hearing impairments. Sign language has the potential to empower a vast amount of the population. With the advancement of computer vision, machine learning, and deep learning techniques, there was plenty of research on American Sign Language (ASL) and British Sign Language (BSL). Despite having so many wide open resources and progress in machine learning, there isn't much thorough research in Bangla Sign Language. There are new possibilities for using Bangla Sign Language and developing a more efficient Bangla Sign Language recognition system. Artificial intelligence algorithms, machine learning models, and multimodal sensor fusion methods have all

been used in existing sign language recognition studies. These studies have produced encouraging results in the recognition and interpretation of sign language motions, but there are still difficulties due to the complexity underlying sign languages, the variety in gestures, and the requirement for performance in real-time.
In the Bangla sign language dictionary [5] there are a total of 36 two-handed Bangla sign alphabets for 51 Bangla written alphabets based on pronunciation. Among these 36 signs 6 of them represent vowels as shown in Fig. 1.1

Figure 1.1: Vowels in Bangla sign language

The rest of the 30 signs represent consonants as shown in the figure-1.2. Also, there are 10 basic signs for numbers 0-9 (fig 1.3). using those alphabetical signs there could be approximately 5000 sets of hand gestures in the Bangla sign language. Often it is very difficult for unsigned people to communicate with signed people because the wide sets of hand gestures are not easy to learn overnight.[2]

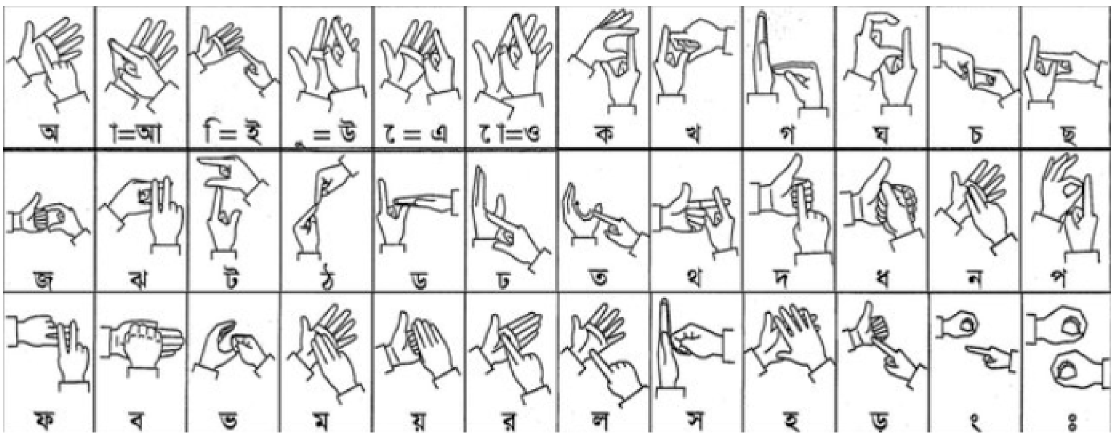


Figure 1.2: Consonants in Bangla sign language

To overcome this problem, so that people can easily communicate with sign language there is a need for a sign language tutor that can teach the complicated hand gesture more accurately as well as give a real-time evaluation experience so that the learners can see the accuracy of their mimicking.

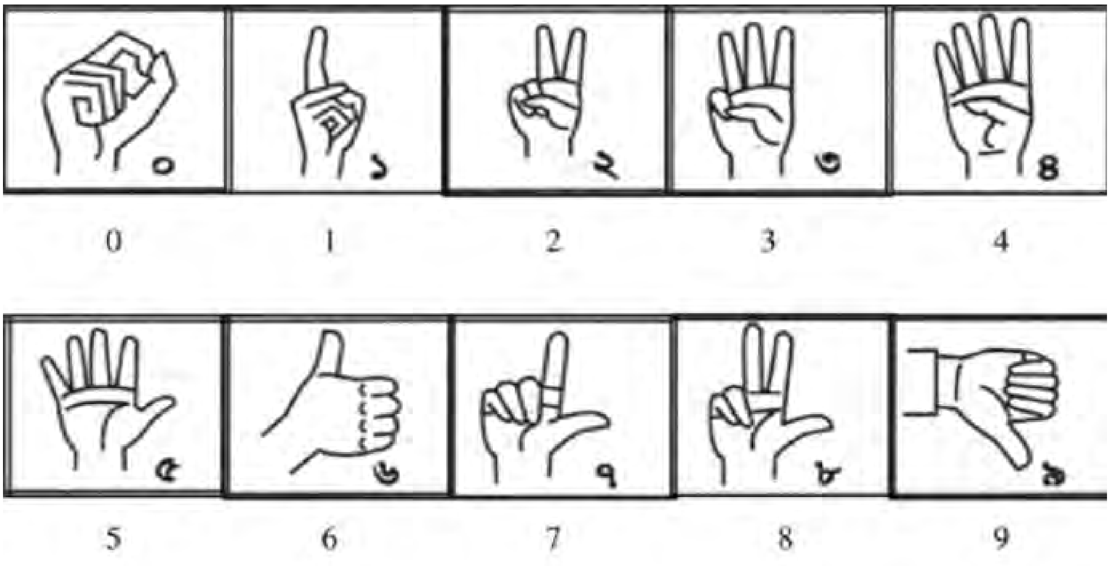


Figure 1.3: Numbers in Bangla sign language

Table 1.1: Statistics of Bangladesh Disability Information System

Gender	Hearing impairment	Mute people
Male	71268	111716
Female	62182	82572
Others	67	106
Total	133517	194394

1.4 Motivation

Picture a world where educational opportunities know no bounds, where every individual, regardless of their hearing abilities, can access quality learning. In a country like Bangladesh, where resources for the Deaf are limited and their progress is impeded, this research addresses a critical need by breaking down communication barriers. A third world, developing country, needs to utilize all the people of the nation to excel as a whole country. Addressing the issue, and incorporating a solution that can enhance the communication life between Bengali people will accelerate the country's growth in many aspects. By integrating the details and nuance of the language into the Avatar Tutor, the project becomes an intermediary as well as a teaching instrument. Communication with the Deaf, mute community, educators, and technology specialists is urged to ensure that it is effective.
In the above table, we can see there are around 327,911 who can make great use of a platform where learning Bangla sign language is cheap and easy. A user-friendly design and firm implementation of an optimized system that teaches Bangla sign language, will bring a life to the communication bridge between those impaired people.

1.5 Research Problem

The primary research problem addressed in our paper "A Smart Avatar Tutor for Mimicking Characters of Bangla Sign Language," depends on the challenges of providing effective and accessible sign language education to the speaking and hearing disabled people in Bangladesh.
As per the statistics from the Department of Social Services, in Bangladesh, there is a significant population with hearing and speaking disability. It reports approximately 194,394 vocal-disabled people and 133,517 hearing-disabled people in Bangladesh [13]. However the use of Bangla sign language in the education system is still limited. It indicated ignorance towards these individual communities. Language and education is a fundamental human right, this gap is necessary to be filled for social integration and personal development. There could be several factors behind the limitation to the accessibility including insufficient funding for the special education system, lack of trained Bangla Sign language instructors, or limited awareness of the needs of those minority people. Those people must learn sign language to communicate and exercise their rights as well as participate in society as responsible citizens.

Also, there have been some technological gaps in Bangla sign language teaching technologies compared to the advancements of other educational technologies. The current technologies focus on the mainstream education system rather than people who need a special education system. Specific needs for sign language imitation make it more challenging to implement. Besides hand gestures, facial expressions, and body language also play a crucial role in sign language learning. It can not be met by basic educational technologies. It needs more precise teaching and evaluation techniques. Capturing the slightest expression difference in facial expressions along with hand movement in real-time is a bit challenging and needs advanced models and algorithms. Addressing this gap may involve the usage of virtual reality, augmented reality, and 3D or 2D modeling.

Sign language teaching tools need to be more interactive and engaging, unlike traditional educational tools. Sign language learning needs more careful observations on the mimicry of gestures in real-time and immediate feedback to increase perfection. So there has been a gap in the need for real-time interactive learning tools. For learners, these tools are not just about learning signs, it is about enhancing their ability to communicate effectively in everyday life. The kind of application that can be accessible from a mobile phone or website online could be helpful and convenient in this case.

There are specific constraints in sign language learning technology, and that is the problem with scalability and customization. The limitations could have several reasons including resource availability, technological constraints, insufficient funding, or specificity of the content. Scalability often needs more resources with intensive and robust data sets. Some tools used in one platform might not be available on others thus creating technological limitations. Different people have different learning paces and techniques and need special care. However, the use of a one-size-fits-all approach could be challenging for individual learning which results in ineffective learning.

There is a lack of robust technology to evaluate the effectiveness and a lack of Bangla sign language learning tools.in any education system, it is important to evaluate the effectiveness of its system. otherwise, there is no point in using it without visible effectiveness. There is always a lack of standardization of sign language proficiency tools or metrics. Without further improvement of the system with time it will no longer remain usable and to do that getting evaluations and feedback are also important and there are not adequate numbers of tools available to do that.

Consequently, our goal is to provide a novel, cutting-edge technological approach to deal with these issues. A smart avatar teacher using the YOLOv8 Nano basic model for real-time sign language interpretations and identification is the suggested solution. With consideration for the particular requirements of Bangladesh's hearing-impaired community, this tool seeks to offer a dynamic, helpful, and easily accessible BSL learning environment. Additionally, the study will concentrate on creating techniques to assess how well this technology helps users communicate in sign language.
Creating a unique, cutting-edge technical answer to such issues is an academic endeavor. A smart avatar teacher that uses the YOLOv8 Nano baseline model for real-time sign language characterization and classification is the suggested solution. With this technology, the hearing-impaired community in Bangladesh will have an easily accessible, engaging, and efficient BSL learning platform that is specifically designed to meet their requirements. The development of techniques to assess this tool's efficacy in enhancing learners' communication using Bangla sign language abilities will be another main focus of the paper. .

Therefore, our research is focused on the noble goal of bridging the communication gap by developing a cost-effective system to teach Bangla Sign Language, thereby facilitating better understanding and interaction. The development of a smart avatar tutor, integrated with the YOLOv8 nano baseline model, effectively enhances the learning of Bangla Sign Language for individuals with hearing impairments in Bangladesh, while ensuring cost efficient system optimization and effective learning.

Examining how well an advanced avatar-based learning platform can meet the specific requirements of the hearing-impaired and mute people community, this research seeks to determine the likelihood that it can be used to teach Bangla Sign Language effectively. It also looks at how the YOLOv8 Nano model may be implemented to enhance the avatar's reliability and adaptability, which will enhance the process of learning as a whole. The inquiry also takes into account the difficulties in optimizing systems to maximize user involvement and educational outcomes.This research will answer the above question by exploring the accuracy obtained by the avatar with the YOLOv8 Nano model to train from the dataset.

1.6 Research Objectives

The goal of the research paper is to address many important goals in developing and implementing a revolutionary educational resource to demonstrate Bangla Sign Language (BSL) to people who are hard of hearing and talking. These goals are intended to guarantee that the application is not just revolutionary in terms of technology but also functions effectively, is available, and is easy to use. The main goals consist of:

- Development of a Smart avatar-based learning system that can engage with learners in an effective and user-friendly way, making it easier for hearing and speaking disabled people.

- Integration of YOLOv8 Nano Baseline Model to optimize for efficiency and accuracy, ensuring that the avatar can work very fast and accurately on low-powered devices in real-time with high precision.

- Enhancing user involvement through interactive learning methods, such as mimicking whatever the avatar shows getting feedback on the accuracy of the mimicry, and adapting the teaching techniques based on education evaluation and performance metrics to improve future learning experience.

- Ensuring accessibility and inclusivity of the system.

- Addressing the gaps in Bangla sign language technology.

- Contribution to assistive educational technology by developing a unique and effective tool for Bangla sign language

By achieving these broad goals, the research paper intends to develop a revolutionary tool for instruction that can improve the communication and learning skills of people with hearing impairments, improving their quality of life and social acceptance.

Chapter 2

Literature Review

2.1 Literature Review

The use of avatars to teach sign language is an approach that has been adopted in numerous research and projects in the past, and increasingly in the present. In the research undertaken by Quandt et al (2020), they used signing avatars created from motion capture, which appeared to have more natural facial expressions, to teach American Sign Language (ASL) in an immersive 3D environment; their system was aptly named SAIL- Signing Avatar and Immersive Learning, which was built-in Unity3D. Emphasis was made on the avatar's hands and eyes. The idea they implemented involved the users physically producing ASL signs taught by the avatar. In other words, they were to mimic the signs demonstrated by the avatar. This resulted in engaging the sensorimotor system of the users, which helped them to learn the signs much quicker. The combination of motion capture, 3D environments, and sensorimotor engagement can be considered a novelty in this field of teaching sign language. Further research is required to obtain more effective outcomes.[5]

Rahaman et al. (2020) applied CNN and LSTM algorithms for sign language recognition in order to train their model in Bangla sign language (BSL). They employed OpenPose to generate a 3D avatar that can replicate Bangla sign language through gestures. The procedure is as follows- first, the user needs to enter their website and select any Bangla number/alphabet that they want to learn. The model will then proceed to demonstrate that character to the user. The user then practices the sign and after measuring their accuracy of performance with cosine similarity, feedback is given back to the user, whether they can progress to the next word or not. Their study established the effectiveness of machine learning in the context of sign language.[12]

Lee et al. (2021) also proposed an ASL learning application prototype, but rather than using an avatar they created the app of a whack-a-mole game with a real-time sign recognition system embedded. To handle static and dynamic datasets they used LSTM RNN with k-Nearest -Neighbour (KNN) method; RNN constituted the Input layer with 30 neurons, the LSTM layer with 28 neurons, the Dense layer with 26 neurons, the Lambda layer with 28 neurons. They achieved an average of 99.44% accuracy rate and 91.82% in 5-fold cross-validation with the use of a leap motion controller.[7]

In their research, Kothadiya et al. (2023) implemented Explainable AI (XAI) in Sign Language Recognition, enhancing the precision of prediction models. They incorporated attention-based ensemble learning, crafting an architecture that utilized ResNet50 with the Self Attention model. This ensemble learning approach exhibited remarkable accuracy, reaching 98.20% specifically on Indian Sign Language (ISL) datasets. Expanding the scope of their work, the researchers applied this methodology to other sign language databases like American Sign Language (ASL) and British Sign Language (BSL). Their findings underscored the superiority of the learning approach coupled with XAI, demonstrating higher accuracy compared to commonly used algorithms such as CNN and RNN. The integration of XAI not only improved accuracy but also provided interpretable insights into the decision-making processes of the model [12].

Efforts have been dedicated to facilitating individuals with hearing impairment in learning to spell words. A notable contribution comes from Savur et al. (2016), who introduced a novel approach employing surface Electromyography (sEMG) to recognize and spell American Sign Language (ASL) letters. The methodology involved comprehensive feature extraction techniques, including time domain analysis, frequency domain analysis, power spectral density, and average power. Subsequent to feature extraction, Principal Component Analysis (PCA) was applied to derive uncorrelated features. Classification methods, specifically Support Vector Machine (SVM) and Ensemble Learning, were then employed to discern and categorize the extracted features, paving the way for effective recognition and spelling of ASL letters[1]

Various approaches exist for sign language recognition, ranging from wearable devices to computer vision and deep learning with neural networks. In a recent study, Zhang, W. (2022), showcased a comprehensive method for extracting hand features from an extensive dataset of images. This involved distinct techniques, including palm-centered apparent feature extraction and feature extraction from wrist joints, utilized in training the model. The culmination of these efforts resulted in an impressive final recognition rate of 91.3%. The study emphasizes that the application of deep image processing technology not only enhances the intelligence of interactions between machines and humans but also holds relevance across a diverse array of fields [17].

In a significant endeavor, Shafiqul et al. (2019) curated an extensive dataset comprising more than 7 thousand samples for the 10 numerical characters and nearly 24 thousand samples representing 35 fundamental Bangla alphabets. This comprehensive dataset served as the input for their Convolutional Neural Network (CNN)-based architecture. The outcome was truly remarkable, with the CNN achieving recognition accuracies of 99.83%, 100%, and 99.80% for basic characters, numerical characters, and character-numerical combinations, respectively. This achievement underscores the effectiveness of their CNN-based model in accurately recognizing and categorizing diverse elements within the Bangla script[3].

Rayeed et al. (2022) employed a dataset comprising static hand gestures representing sign digits in Bangla Sign Language (BdSL). The researchers utilized various classifiers, including Support Vector Machine (SVM), XGBoost (XGB), and k-nearest Neighbors (KNN), to classify the dataset. The depth values of key hand points were subsequently estimated using MediaPipe.[2] On a parallel note, Mehedi et al. (2020) opted for the Ishara-Lipi dataset in their exploration of a novel Convolutional Neural Network (CNN). They conducted feature extraction to capture crucial elements for accurate alphabet recognition within the dataset. Leveraging the Keras framework and TensorFlow, their model achieved an impressive accuracy rate of 99.22%. These distinct approaches showcase the versatility and effectiveness of various methodologies in the realm of sign language gesture recognition[9].

In the comprehensive exploration conducted by Kumar et al. (2023), the strategic amalgamation of the Convolutional Neural Network (CNN) and the Transfer Learning method proved instrumental in attaining a noteworthy accuracy of 90.1%. This robust combination not only showcased the prowess of advanced neural network architectures but also demonstrated the adaptability of transfer learning techniques in optimizing model performance.
Moreover, the research team's decision to employ Python and OpenCV for sign detection further enriched the study's methodological diversity. The utilization of these widely embraced programming tools not only facilitated efficient implementation but also emphasized the practical applicability of their approach in real-world sign language recognition scenarios. This multifaceted strategy not only contributes to the advancement of sign language research but also underscores the importance of leveraging a spectrum of technologies for holistic and effective solutions in this domain[14].

Rahaman et al.'s (2018) groundbreaking research not only presented a two-phase modeling algorithm for automatic recognition but also showcased a meticulous approach that significantly contributed to advancing the field of Bangla Sign Language recognition. The intricately designed first phase, dedicated to hand-sign classification, laid the foundation for the subsequent phase, where the focus shifted to recognizing specific hand-signs. Through the adept use of a two-step classifier, the research team achieved not only an impressive mean accuracy of 95.83% but also surpassed the 90% accuracy threshold in detecting hand-sign-spelled words, composite numerals, and sentences. This nuanced and methodical approach highlights not only the algorithm's effectiveness but also its versatility across diverse facets of Bangla Sign Language recognition, making it a noteworthy contribution to the evolving landscape of sign language research [2].

In the realm of Bangla Sign Language (BdSL) research, Islam et al. (2019) employed Convolutional Neural Network (CNN) technology to decode static hand gestures. Their CNN classification, featuring four convolutional layers, achieved an accuracy of 88% in recognizing and categorizing these gestures, marking a significant stride in the field of sign language interpretation. Additionally, Basnin et al. proposed an innovative approach by integrating CNN and Long Short-Term Memory (LSTM) models. Their hybrid model exhibited commendable performance, boasting a training accuracy of 90% and an 88.7% testing accuracy. This integration of CNN and LSTM highlights a holistic strategy that considers both spatial and temporal aspects, showcasing the potential for nuanced and accurate sign language recognition. The diverse methodologies employed by these research teams contribute valuable insights to the broader landscape of assistive technologies, promising advancements in communication accessibility for individuals using Bangla Sign Language and beyond [4].

In this study, the detection of traffic signals was tested with 3 models- namely YoloV5, YoloV7, and YoloV8. The motivation for this research was that in the present time, there is a growth in autonomous vehicles. And one very big obstacle for this industry is reducing the amount of accidents. As a result, being able to detect the traffic lights is crucial and this study kept the same parameters for the 3 models and analyzed values such as prediction, recall, mAP, complexity matrix, and F1 score. In the results obtained, YoloV8 gave 14% and 13% mAP50-95 improvement compared to the other models. [15]

This paper discussed all the different versions of the YOLO model (from v2 to v8). In the present, real-time object detection has allowed researchers to make breakthroughs in many different fields. Due to its high speed and accuracy, the YOLO model is well-suited for such tasks. In general, Yolo works by dividing the image into grids and creating bounding boxes in each grid. Next, it predicts the classes of objects for each of the bounding boxes. Redundant bounding boxes are removed while the confident boxes are kept, this process is known as Non-Maximum Suppression (NMS). Due to v8 being anchor-free, it has fewer bounding boxes, thus less time taken during NMS. Hence v8 yields the most efficient results thus far. [16]

This study focuses on the effectiveness of using Yolo models for forest fire detection. Yolov5, v7, and v8 models were first designed and then trained for 50 and 100 epochs. For training the images from the dataset were pre-processed. Mean Average Precision(mAP), Precision, Recall, and F1 score were obtained and compared for each model. Evaluating these scores it was seen that overall Yolov8n performed better than the other models; it reached precision and recall scores of over 90% and registered the highest mAP95(%). However, its training time was also significantly longer. [10]

Chapter 3

Methodology

The aim is to create a learning platform to teach Bangla Sign language with the help of an avatar, as mentioned in the previous chapter.

3.1 Work Plan

Figure 3.1: The flowchart of our work plan.

Following the collection of the dataset, preprocessing techniques (Histogram Equalization, Noise reduction) were applied. The data will be split into an 8:2 ratio, i.e. 80% for training and 20% for testing. The models used for training is YOLO V8. Evaluation will be made using accuracy, precision, recall, and F-Score. Finally, the model will be tested in real-time sign language.

3.2 Dataset

Obtaining a dataset for Bangla sign language was a challenging task. Popular sign languages such as ASL have more available datasets. After a lot of searching, we found a dataset that is free to use from the website ScienceDirect [11]. The dataset consists of a total of 29,428 images of hand gestures for numbers and alphabets in Bangla.

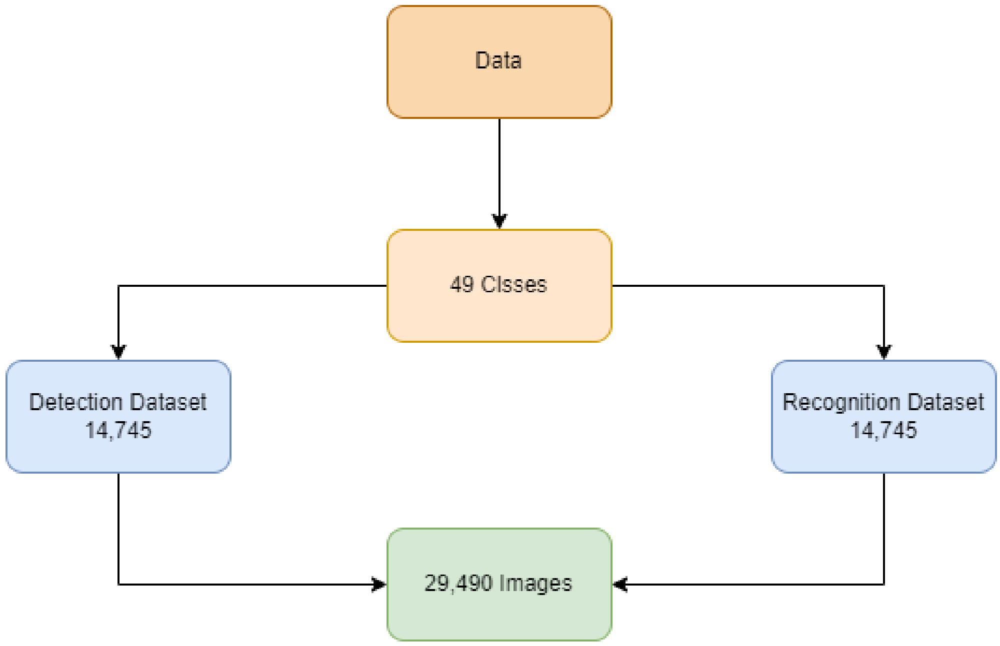


Figure 3.2: Dataset contents

Table 3.1: Dataset class distributions

Label No.	Class Name	Total No. of images
0	অ	600
1	আ	600
2	ই	600
3	উ	600
4	উ	600
5	ও	600
6	ক	600
7	খ	600
8	গ	600
9	ঘ	600
10	চ	600
11	ছ	600
12	জ	600
13	ঝ	600
14	ট	600
15	ঠ	600
16	ড	600
17	ঢ	600
18	ত	600
19	থ	600
20	দ	600

Table 3.1 shows the distribution of the preliminary dataset obtained for the first 20 characters. While the database is freely available, it still requires pre-processing.

3.3 Pre-processing

Pre-processing:

i) Images from the dataset were converted to greyscale.

ii) The images were trimmed, and the background was blurred to help in the detection of the hand.

iii) Applied Histogram Equalization and Noise Reduction.

iv) Finally the images are stored in a new dataset.

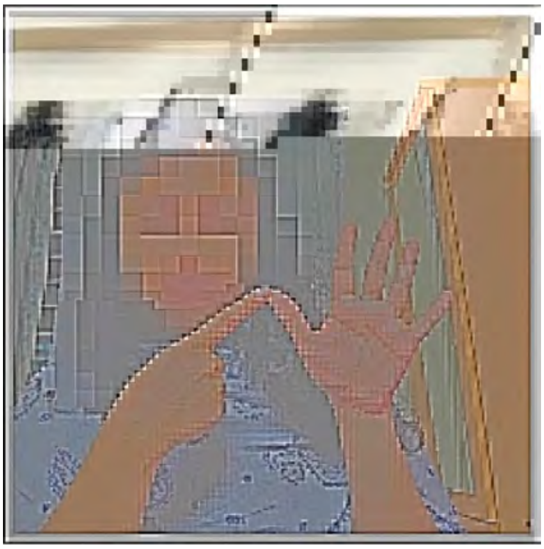


Figure 3.3: Image before preprocessing

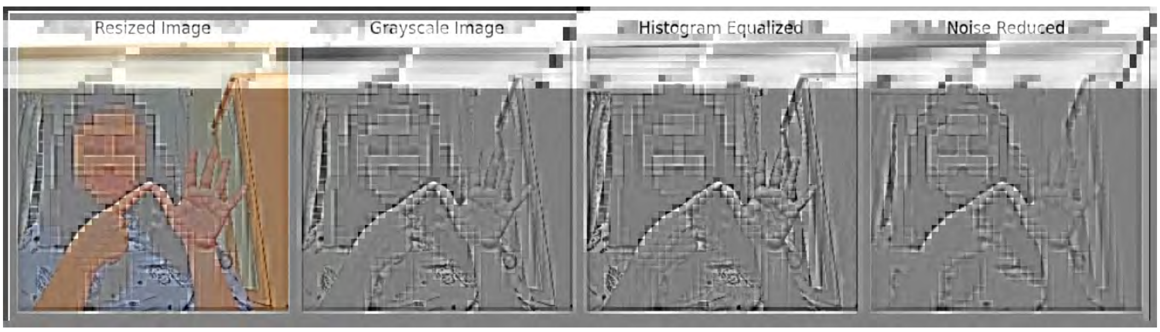


Figure 3.4: Image after preprocessing

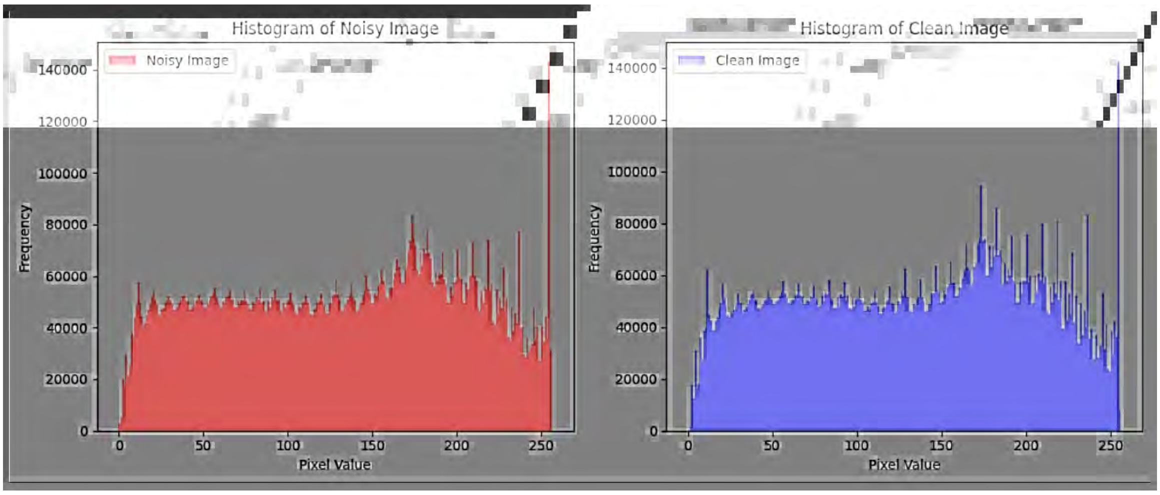


Figure 3.5: Histograms of Noisy vs Clean Image

3.4 Proposed Method

3.4.1 Yolo V8

YOLO, an acronym for "You Only Look Once," represents a breakthrough in object detection algorithms, acclaimed for its speed and efficiency. Unlike traditional models that process parts of an image sequentially, YOLO applies a singular convolutional neural network (CNN) to the entire image simultaneously. This approach divides the image into a grid, with each cell responsible for predicting bounding boxes and associated probabilities. For detecting multiple real-time objects, it provides optimal results than most other models.

The essence of YOLO's architecture is grounded in CNNs, which extract features from the input image. These features are pivotal for the prediction of bounding boxes and class probabilities. The bounding boxes encompass objects and are accompanied by confidence scores indicating the precision of the prediction. The model concurrently predicts the probability of each object class within these boxes.

A paper published September 13, 2023, evaluated the performances of YOLO V5, YOLO V7 and YOLO V8 on various traffic sign detection. The researchers trained YOLO V5, YOLO V7, and YOLO V8 models using a dataset containing 4650 photographs of 15 different traffic signs. They analyzed various performance metrics such as prediction, recall, mAP (mean Average Precision), complexity matrix, and F1 score. The results showed that the YOLO V8 model performed better than the others, with a 14% and 13% improvement in mAP50-95 over YOLO V5 and YOLO V7, respectively. [15]
Overall, the YOLO V8 seems to provide the most optimal performance in sign detection. Therefore, our research is mainly conducted based on the YOLO V8 model.

- An input layer that processes the image and divides it into a grid.
- Several convolutional layers for feature extraction.

- Bounding box prediction for each grid cell.
- Class probability prediction for detected objects.

Mathematical Formulations: The YOLO architecture incorporates several key mathematical concepts:
Bounding Box Prediction: Each grid cell predicts bounding boxes with the following formulas:

$$\begin{aligned}
B_x &= \sigma(t_x) + c_x \\
B_y &= \sigma(t_y) + c_y \\
B_w &= p_w e^{t_w} \\
B_h &= p_h e^{t_h}
\end{aligned}$$

where (B_x, B_y) and (B_w, B_h) are the center coordinates and dimensions of the box, (t_x, t_y, t_w, t_h) are the network outputs, and (c_x, c_y) are the top-left coordinates of the grid cell.
Confidence Score: The confidence score for each bounding box is:

$$\text{Confidence} = \Pr(\text{Object}) \times \text{IOU}_{\text{pred}}^{\text{truth}}$$

where IOU is the Intersection Over Union between the predicted box and the ground truth.
Class Probabilities: The conditional class probability is given by:

$$\Pr(\text{Class}_i|\text{Object}) = \frac{e^{(c_i)}}{\sum_j e^{(c_j)}}$$

where c_i is the class prediction for class i.

Loss Function: The loss function in YOLO is a composite of:

- Coordinate loss for bounding box accuracy.
- Confidence loss for objectness prediction.
- Classification loss for correct class prediction.

Improvements in New Versions: Each new version of YOLO usually brings improvements such as more efficient feature extraction, better handling of varied object sizes, and improved accuracy and speed. and Yolo V8 is an anchor-free model. This means it predicts directly the center of an object instead of the offset from a known anchor box.

3.4.2 Xception

Xception, an abbreviation for "Extreme Inception," is a deep learning neural network architecture that modifies and extends the Inception architecture. It was introduced by François Chollet and is known for its efficiency in image recognition tasks.
Architecture: The Xception architecture is built on depth-wise separable convolutions, offering a more efficient way to learn spatial hierarchies.

- *Depthwise Separable Convolutions:* Splits the convolution process into depth-wise and point-wise convolutions, reducing parameters and computations.

- *Entry Flow:* Consists of standard convolutions and max-pooling layers, processing the input image.

- *Middle Flow:* The core of the network, with a series of depth-wise separable convolutions for complex feature extraction.

- *Exit Flow:* Prepares the final feature map with separable convolutions and fully connected layers.

Mathematical Formulas: Depthwise separable convolution, the key concept in Xception, involves:

1. *Depthwise Convolution:* For input tensor X and depthwise convolution kernel K_d:
$$Y_{k,i,j} = \sum_{m,n} X_{k,i+m,j+n} \cdot K_{d,k,m,n}$$

2. *Pointwise Convolution:* Applied after depthwise convolution with a 1x1 kernel K_p:
$$Z_{l,i,j} = \sum_{k} Y_{k,i,j} \cdot K_{p,l,k}$$

Loss Function and Optimization: Xception uses categorical cross-entropy for classification tasks and is optimized using algorithms like Adam or SGD.
Applications: It excels in image classification and object detection, often outperforming other architectures in accuracy and efficiency.
Conclusion: Xception's efficient use of model parameters through depthwise separable convolutions makes it a powerful tool in computer vision tasks, balancing accuracy and computational efficiency.

3.4.3 InceptionV3

InceptionV3 is an evolution of the original Inception architecture, part of the GoogLeNet series, designed for improved accuracy in image classification while maintaining computational efficiency.
Architecture: InceptionV3 introduces several innovations and optimizations:

- *Factorized Convolutions:* Reduces the number of connections by breaking down larger convolutions into smaller ones.

- *Asymmetric Convolutions:* Uses a combination of 1xN and Nx1 convolutions instead of larger ones to reduce computational cost.

- *Auxiliary Classifiers:* Adds secondary loss during training to combat the vanishing gradient problem.

- *Grid Size Reduction:* Efficiently reduces grid sizes to increase the depth and width of the network without a significant rise in computational cost.

- *Batch Normalization:* Used extensively to accelerate training.

Mathematical Formulas: Key concepts include:

1. *Factorized Convolution:* For a kernel size of NxN, it is factorized into two convolutions of sizes 1xN and Nx1, reducing parameters from N^2 to $N + N$.
2. *Auxiliary Classifier Loss:* The total loss L_{total} is a combination of the main classifier loss L_{main} and the auxiliary loss L_{aux}:

$$L_{total} = L_{main} + \lambda \cdot L_{aux}$$

where λ is a weighting factor.

Applications: InceptionV3 is used in various image recognition tasks, including object detection and classification.

Conclusion:
InceptionV3 is notable for its use of factorized and asymmetric convolutions, auxiliary classifiers, and efficient grid size reduction, making it a powerful model for image recognition tasks.

3.4.4 RestNet50V2

ResNet50V2 is an improved version of the original ResNet50, part of the Residual Networks family, known for its efficiency in image classification tasks.
Architecture: ResNet50V2 is characterized by its deep architecture and use of residual connections:

- *Bottleneck Architecture:* Consists of three layers in each block, with 1x1 and 3x3 convolutions.
- *Identity Shortcut Connections:* Skip connections that perform identity mapping to mitigate the vanishing gradient problem.
- *Pre-activation:* Batch normalization and ReLU activation are performed before convolutions in each block.

Mathematical Formulas: Key concepts include:

1. *Residual Function:* For input x, the output y is given by:

$$y = F(x, \{W_i\}) + x$$

where $F(x, \{W_i\})$ is the residual mapping.

2. *Pre-activation:* The function can be represented as:

$$F = W_2 \cdot \text{ReLU}(\text{BN}(W_1 \cdot \text{ReLU}(\text{BN}(x))))$$

where W_1 and W_2 are the weights of the convolutions.

Applications: ResNet50V2 is used in image classification, object detection, and other computer vision tasks.
Conclusion: ResNet50V2's architecture, with its deep layers and residual connections, makes it a powerful model for complex image recognition tasks.

3.5 System Design

This is the system design for the smart avatar tutor.

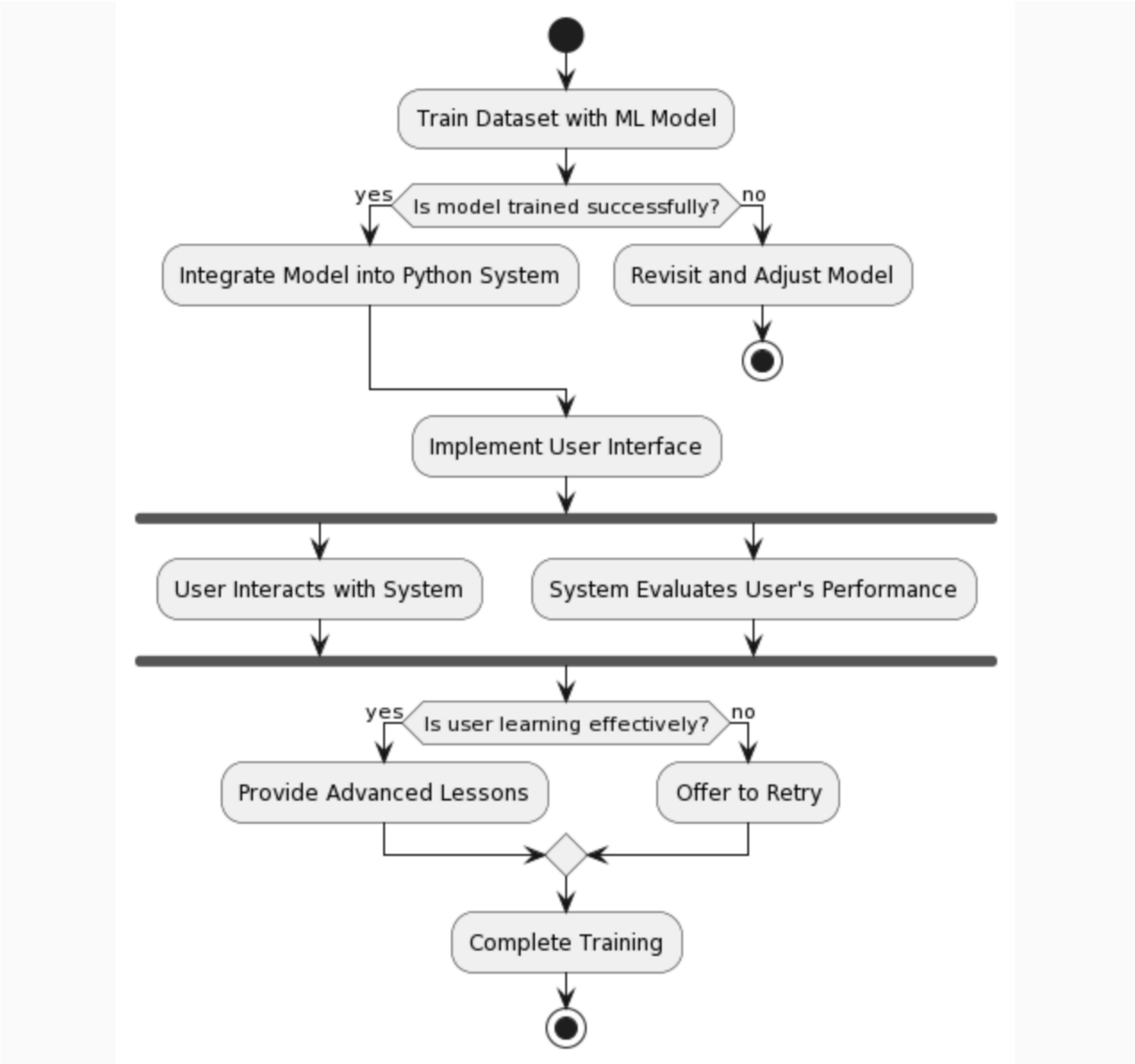


Figure 3.6: Activity Diagram of Smart Avatar Tutor

This is the system design for the smart avatar tutor.

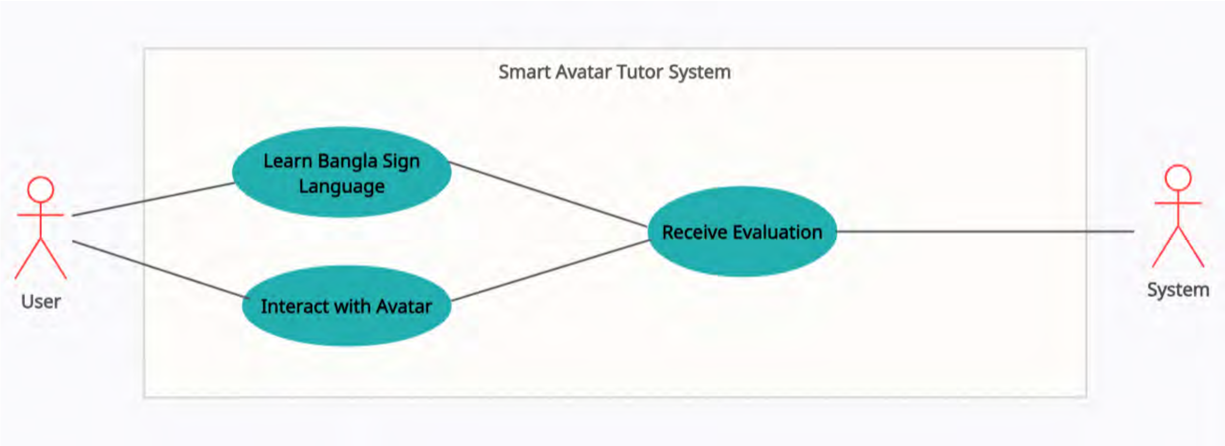


Figure 3.7: Use Case of Smart Avatar Tutor

3.6 System User Interface and Experience

First, the user needs to log in to our website. They will have the option to select a Bangla number, alphabet, or word. Next, the avatar will prompt and demonstrate the sign of the selected option step-by-step. The user can opt for a replay from the avatar. Then the user can practice the sign, which will be captured. If the accuracy of their performance is above a certain threshold, the user will be prompted to learn a new word/alphabet, else will be prompted to practice it again.

As the user practices the sign, their motion is captured through the camera. Still images are created into a channel from where we detect the hand and apply the steps described in the pre-processing. Then points are added to determine what sign was being performed by the user. Lastly, we use cosine similarity to compare between the actual sign and the one done by the user.

We have designed an initial Figma design for a user-friendly interface for enriching the learner experience. Below are the screens for the system.

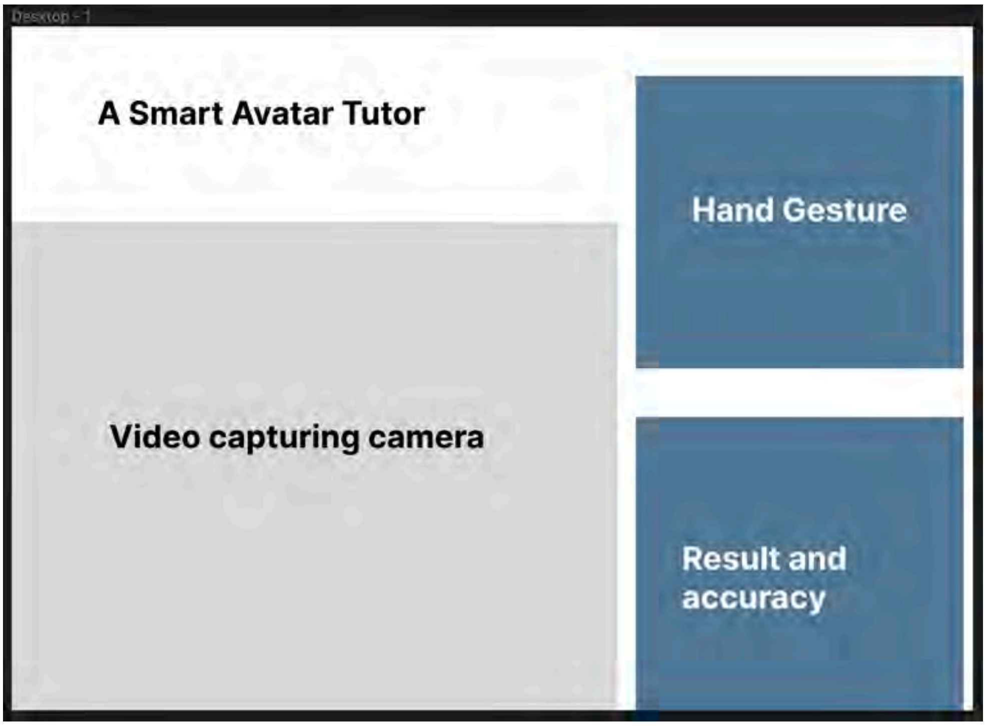


Figure 3.8: Welcome Page Interface

The interface when the user runs the system. The color and combination are selected based on the standard industry interface. The design is to make the software more user friendly. In the wireframe, the big portion of the screen will carry camera screen where user will able to record or see himself mimicking Bangla characters. On the bottom right side, he will see accuracy of the mimicked character. On the top right side, the user will see the hand gesture the user wants to mimick.

Chapter 4

Implementation and Results

This section aims to describe the implementation of the proposed models for training a model Bangla sign language. The implementation of the model consists of three stages. Firstly, input data preprocessing. Secondly, labeling the dataset based on the Bangla language characters. Thirdly, model training and testing the model for use.

4.1 Implementation

The proposed models for building a model that YOLO V8. The dataset. YOLO (You Only Look Once) is an object detection algorithm that can be adapted for real-time Bangla Sign Language detection. YOLOv8, an improved version of YOLO, offers better accuracy and performance. Preprocess the dataset by resizing images or video frames to the input size expected by YOLOv8. Normalize the pixel values if necessary. YOLOv8 predicts the center of an object directly, unlike previous versions that relied on anchor boxes. This approach simplifies the detection process and speeds up Non-Maximum Suppression (NMS). Changes in the convolution layers, including replacing the 6x6 conv with a 3x3 in the stem and modifications in the main building block. An image augmentation technique is used during training, which is turned off in the last ten epochs to improve performance. YOLOv8 shows high accuracy in benchmarks like Microsoft COCO and Roboflow 100, outperforming previous models like YOLOv5 and YOLOv7. YOLOv8 NanoBaseline will be used to train the dataset.

```
class DetectionThread(QThread):
    change_pixmap_signal = pyqtSignal(bytes)
    character_signal = pyqtSignal(int)
    recognition_accuracy = pyqtSignal(float)
    next_character = pyqtSignal(int)
    next_character_ready = pyqtSignal()

    def __init__(self, model_path, parent=None):
        super().__init__(parent)
        self.model_path = model_path
        self.stopped = False
        self.lst_character = []
        self.count = 0
        self.char = 1

    def run(self):
        cap = cv2.VideoCapture(0)
        yolov8_detector = YOLOv8(self.model_path, conf_thres=0.6, iou_thres=0.5)

        while not self.stopped:
            ret, frame = cap.read()
            if ret:
                boxes, scores, class_ids = yolov8_detector(frame)
                if len(class_ids) > 0:
```

Figure 4.1: Code Implementation of System

The below images shows the process to convert normal image to avatar tutor hand gesture for mimicking the sign langauge.

Figure 4.2: Actual Image of Character Sign

Figure 4.3: Rotated and Background Removed of the Actual Sign

Figure 4.4: 3D Avatar Converted Sign

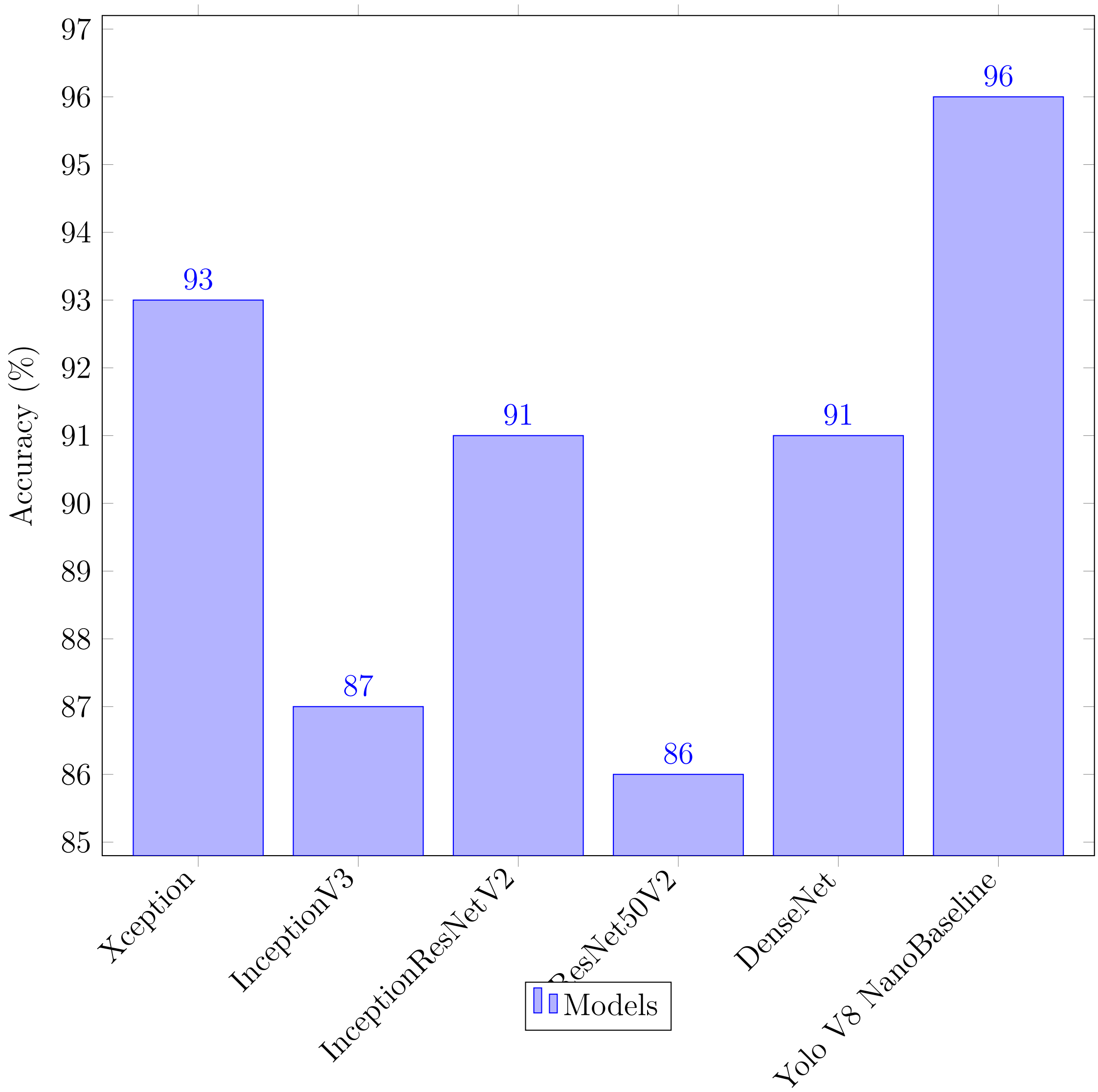
In our research paper, we developed a process to convert hand gestures into a 3D avatar tutor sign. Initially, we captured high-resolution images of various hand gestures. These images underwent several image manipulation techniques, including background removal, rotation, and resizing to ensure uniformity and clarity. Subsequently, we employed illustration techniques to transform these processed 2D images into 3D avatars. This involved edge detection, depth mapping, and 3D modeling to create realistic, textured avatars. Finally, we integrated these avatars with YOLO V8 in the system for real-time gesture recognition, allowing the users to mimic the identified gestures accurately. The entire system was rigorously tested and optimized for responsiveness and accuracy, demonstrating an effective method for creating interactive 3D avatar tutors from hand gestures.

4.2 Results

4.2.1 Model Analysis

Researching and implementing some of the pre-trained models using the dataset we found out that Xception is 93%, InceptionV3 is 87%, InceptionRestNetV2 91%, RestNet50V2 86% and YOLO V8 is 96% accurate. [11]

Comparison of Different Models' Accuracy



The following table shows the performance metrics of the object detection model after training on various classes:

Table 4.1: Training Results of Each Characters

Label	Character	Images	Box(P)	R	mAP50	mAP50-95
0	অ	2940	0.789	0.933	0.848	0.784
1	আ	2940	0.984	1	0.995	0.908
2	ই	2940	1	1	0.995	0.943
3	উ	2940	0.866	0.967	0.959	0.908
4	এ	2940	1	0.833	0.917	0.880
5	ও	2940	0.952	1	0.994	0.899
6	ক	2940	0.982	0.917	0.956	0.860
7	খ	2940	1	0.983	0.991	0.843
8	গ	2940	1	1	0.995	0.893
9	ঘ	2940	0.981	0.883	0.940	0.905
10	চ	2940	1	1	0.995	0.786
11	ছ	2940	0.966	0.950	0.974	0.898
12	জ	2940	1	1	0.995	0.882
13	ঝ	2940	1	0.927	0.963	0.788
14	ট	2940	0.937	0.983	0.989	0.943
15	ঠ	2940	1	0.900	0.950	0.829
16	ড	2940	1	1	0.995	0.892
17	ঢ	2940	1	1	0.995	0.896
18	ত	2940	1	0.750	0.875	0.839
19	থ	2940	0.983	0.983	0.991	0.883
20	দ	2940	0.984	1	0.995	0.897
21	ধ	2940	1	1	0.995	0.885
22	প	2940	1	0.967	0.983	0.890
23	ফ	2940	1	0.950	0.975	0.770
24	ব	2940	1	0.933	0.967	0.861
25	ভ	2940	1	1	0.995	0.796
26	ম	2940	0.983	0.967	0.976	0.863
27	য়	2940	1	0.967	0.983	0.840
28	র	2940	0.951	0.650	0.811	0.705
29	ল	2940	0.946	0.883	0.931	0.840
30	ন	2940	0.939	0.767	0.865	0.797
31	স	2940	0.983	0.983	0.989	0.929
32	হ	2940	0.983	0.983	0.982	0.943
33	ড়	2940	1	0.833	0.917	0.836
34	ং	2940	1	1	0.995	0.918
35	ঃ	2940	1	0.983	0.991	0.835
36	০	2940	1	0.983	0.991	0.858
37	১	2940	1	1	0.995	0.944
38	২	2940	0.984	1	0.995	0.924
39	৩	2940	0.896	1	0.988	0.957
40	৪	2940	1	0.883	0.942	0.897
41	৫	2940	0.984	1	0.995	0.970
42	৬	2940	1	1	0.995	0.920
43	৭	2940	1	1	0.995	0.934

Continued on next page

Table 4.1 continued from previous page

Label	Character	Images	Box(Precision)	Recall	mAP50	mAP50-95
44	৮	2940	1	0.900	0.950	0.922
45	৯	2940	1	0.950	0.975	0.788
46	্	2940	0.983	1	0.982	0.800
47	space	2940	1	0.983	0.991	0.828
48	ঞ	2940	1	1	0.995	0.826

The provided data represents the performance metrics of an object detection model trained to recognize various Bangla characters and numerals. The evaluation was conducted over 2940 images, encompassing 2939 instances of different classes. The key performance indicators used are Precision (Box P), Recall (R), mean Average Precision at 50% IoU threshold (mAP50), and mean Average Precision across IoU thresholds from 50

- **General Accuracy**: The model achieved high overall precision (0.981) and recall (0.951), indicating strong detection capabilities across all classes.
- **mAP50**: The mean Average Precision at 50% IoU (Intersection over Union) is 0.968, suggesting that the model is highly accurate in detecting objects with at least 50% overlap with ground truth.
- **mAP50-95**: The mean Average Precision across a range of IoU thresholds (0.87) demonstrates the model's robustness and accuracy in various conditions.

The following table shows the performance metrics of the object detection model validation on various classes:

Table 4.2: Validation Results of Each Characters

Label	Character	Images	Box(P)	R	mAP50	mAP50-95
0	অ	2940	0.789	0.933	0.852	0.790
1	আ	2940	0.984	1	0.995	0.918
2	ই	2940	1	0.983	0.991	0.940
3	উ	2940	0.879	0.967	0.961	0.909
4	এ	2940	1	0.833	0.917	0.880
5	ও	2940	0.968	1	0.994	0.902
6	ক	2940	0.982	0.917	0.956	0.855
7	খ	2940	1	0.983	0.991	0.843
8	গ	2940	1	1	0.995	0.897
9	ঘ	2940	0.981	0.883	0.940	0.909
10	চ	2940	1	1	0.995	0.787
11	ছ	2940	0.966	0.950	0.974	0.896
12	জ	2940	1	1	0.995	0.889
13	ঝ	2940	1	0.927	0.963	0.788
14	ট	2940	0.937	0.983	0.989	0.942
15	ঠ	2940	1	0.900	0.950	0.828
16	ড	2940	1	1	0.995	0.895
17	ঢ	2940	1	1	0.995	0.895

Continued on next page

Table 4.2 continued from previous page

Label	Character	Images	Box(P)	R	mAP50	mAP50-95
18	ত	2940	1	0.750	0.875	0.839
19	থ	2940	0.983	0.983	0.991	0.885
20	দ	2940	0.984	1	0.995	0.900
21	ধ	2940	1	1	0.995	0.882
22	প	2940	1	0.967	0.983	0.890
23	ফ	2940	1	0.950	0.975	0.769
24	ব	2940	1	0.933	0.967	0.861
25	ভ	2940	1	1	0.995	0.799
26	ম	2940	0.983	0.967	0.976	0.862
27	য়	2940	1	0.967	0.983	0.836
28	র	2940	0.951	0.650	0.811	0.705
29	ল	2940	0.946	0.883	0.931	0.843
30	ন	2940	0.939	0.767	0.865	0.799
31	স	2940	0.983	0.983	0.990	0.929
32	হ	2940	0.983	0.983	0.981	0.944
33	ড়	2940	1	0.833	0.917	0.836
34	ং	2940	1	1	0.995	0.918
35	ঃ	2940	1	0.983	0.991	0.835
36	০	2940	1	0.983	0.991	0.855
37	১	2940	1	1	0.995	0.942
38	২	2940	0.984	1	0.995	0.922
39	৩	2940	0.896	1	0.990	0.961
40	৪	2940	1	0.883	0.942	0.895
41	৫	2940	0.984	1	0.995	0.969
42	৬	2940	1	1	0.995	0.922
43	৭	2940	1	1	0.995	0.936
44	৮	2940	1	0.900	0.950	0.923
45	৯	2940	1	0.950	0.975	0.792
46	্	2940	0.983	1	0.982	0.803
47	space	2940	1	0.983	0.991	0.828
48	ঞ	2940	1	1	0.995	0.820

The below are the result of detection of images while training the model. It was able to perfectly detect most hand gestures accurately.

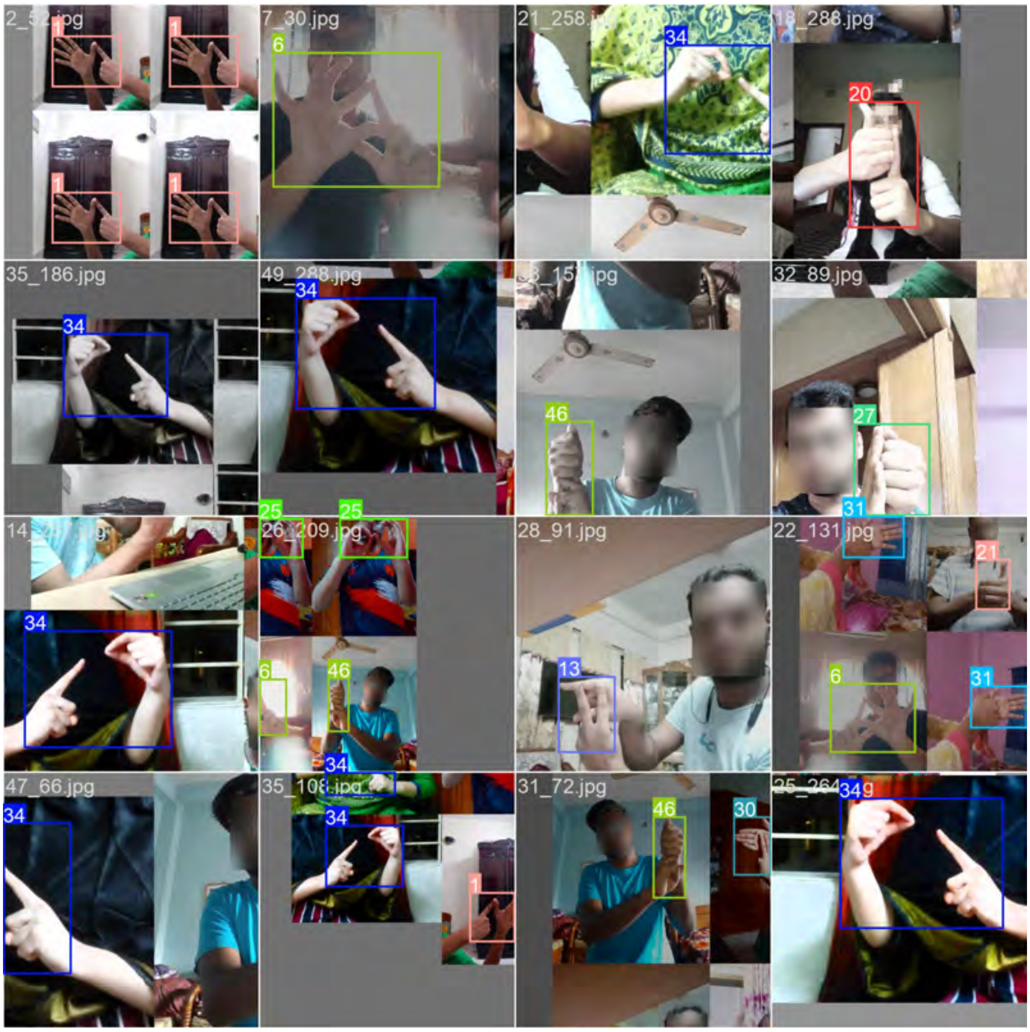


Figure 4.5: Yolo V8 Training batch 0

Figure 4.6: Yolo V8 Validation Batch 0 Labels

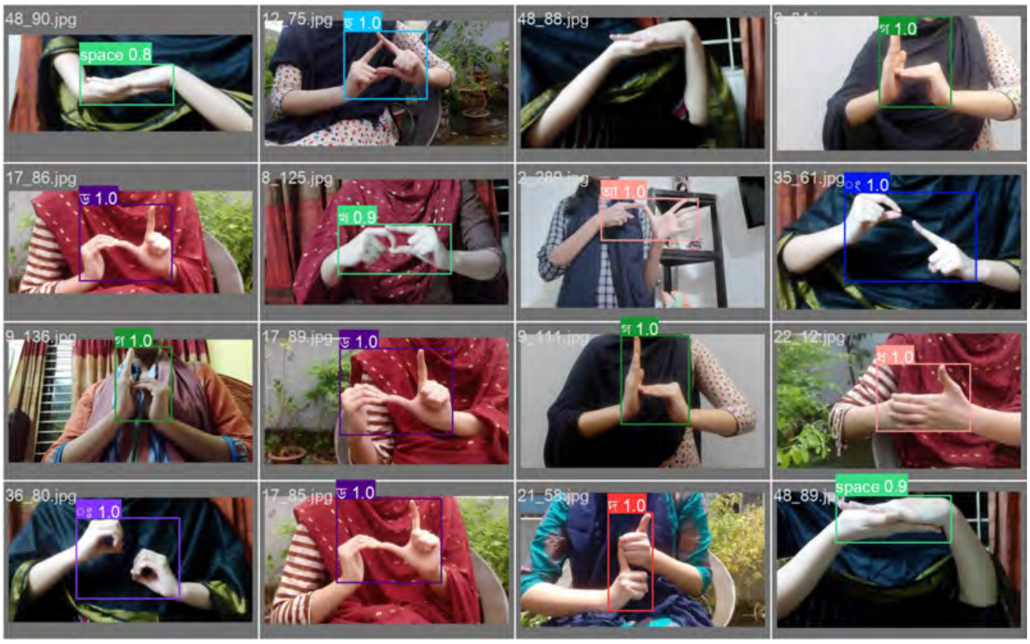


Figure 4.7: Yolo V8 Validation Batch 0 Prediction

Here is the confusion matrix from the training result.

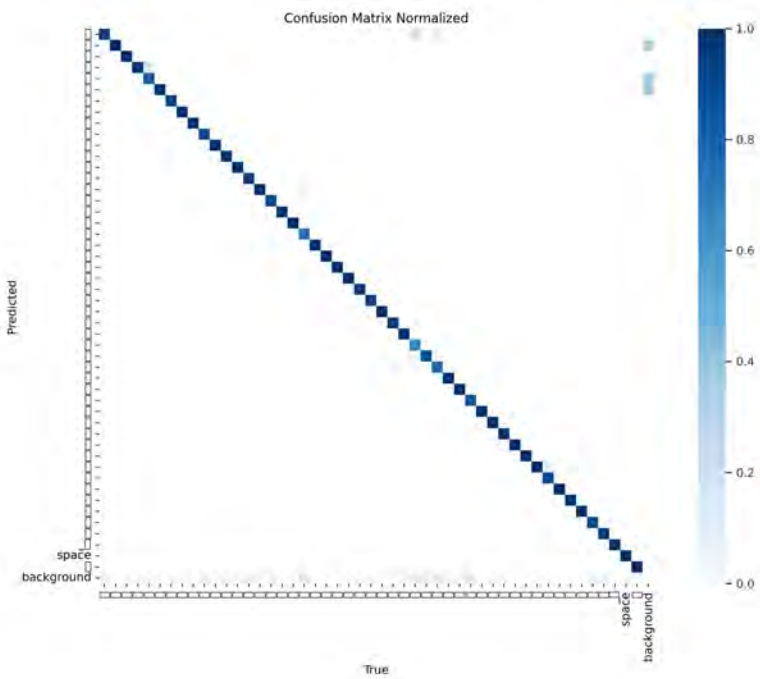


Figure 4.8: Confusion Matrix Normalized

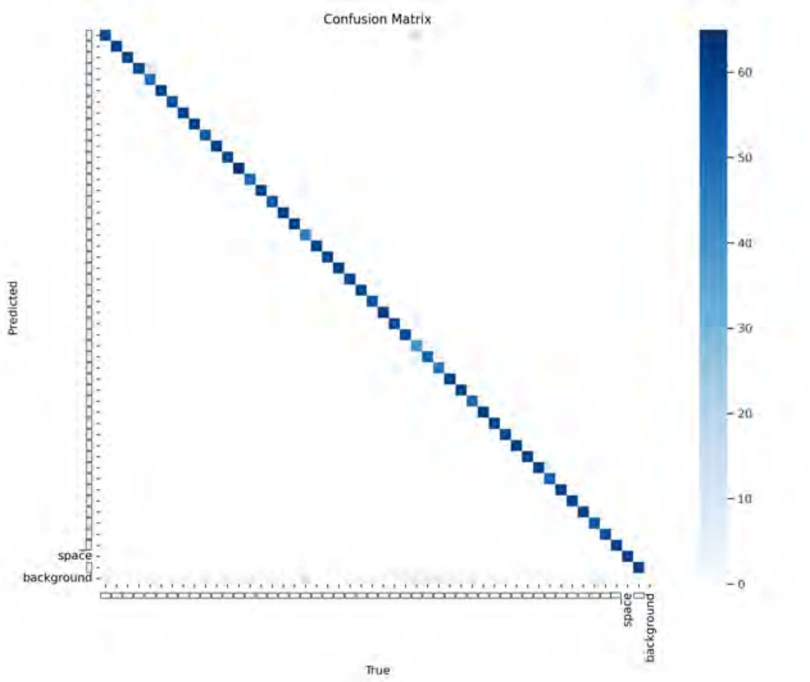


Figure 4.9: Confusion Matrix

The formula for confusion matrix

Predicted		**Actual**
	Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

Table 4.3: Confusion Matrix

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$
$$\text{Precision} = \frac{TP}{TP + FP}$$
$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN}$$
$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$
$$\text{Specificity} = \frac{TN}{TN + FP}$$

Here we can see the precision-recall curve

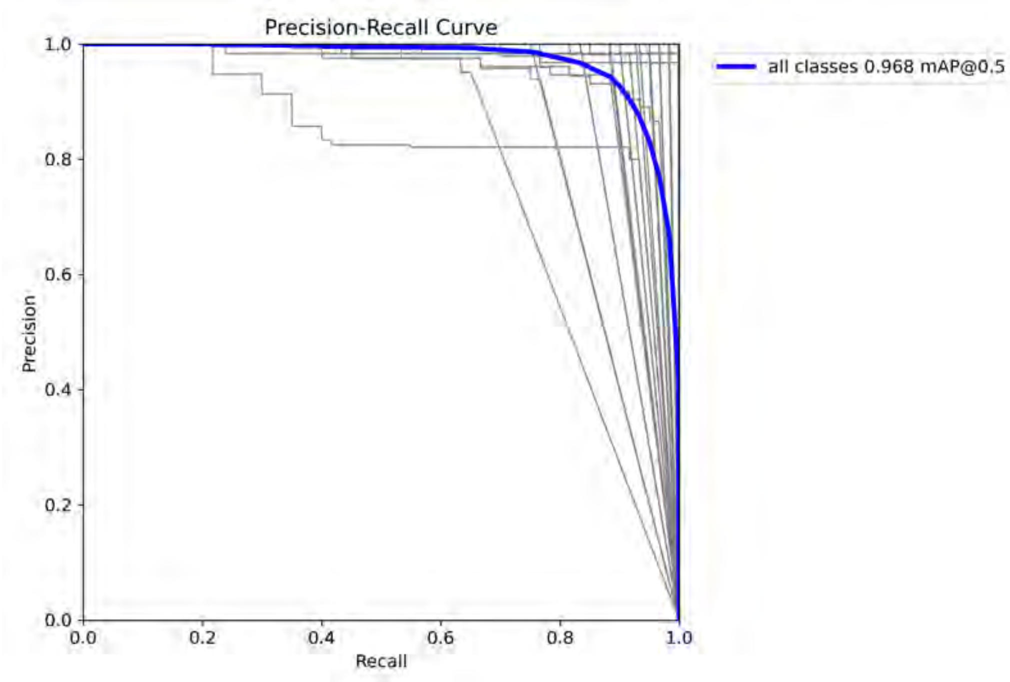


Figure 4.10: Precision Recall Curve

Here we can see the precision curve

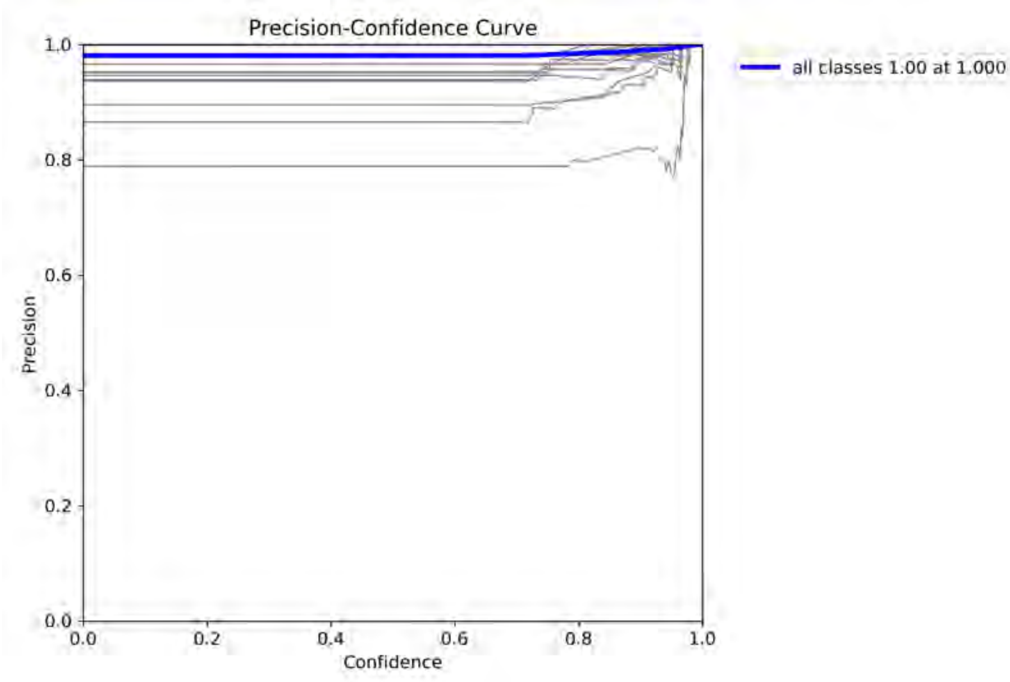


Figure 4.11: Precision-Confidence Curve

Here we can see the recall confidence curve

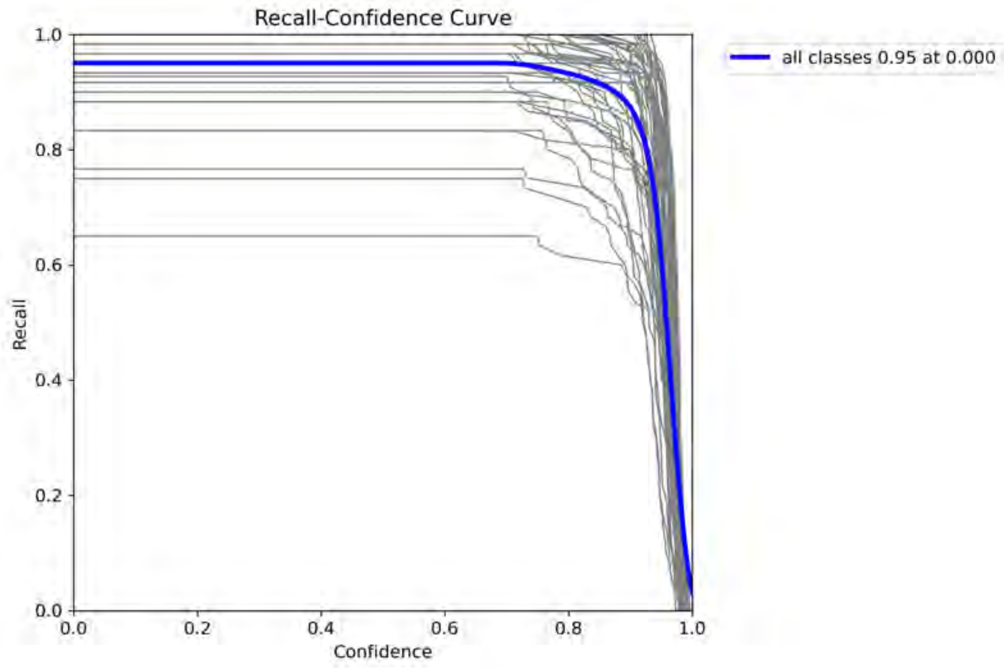


Figure 4.12: Recall Confidence Curve

Here is the confusion matrix for validation.

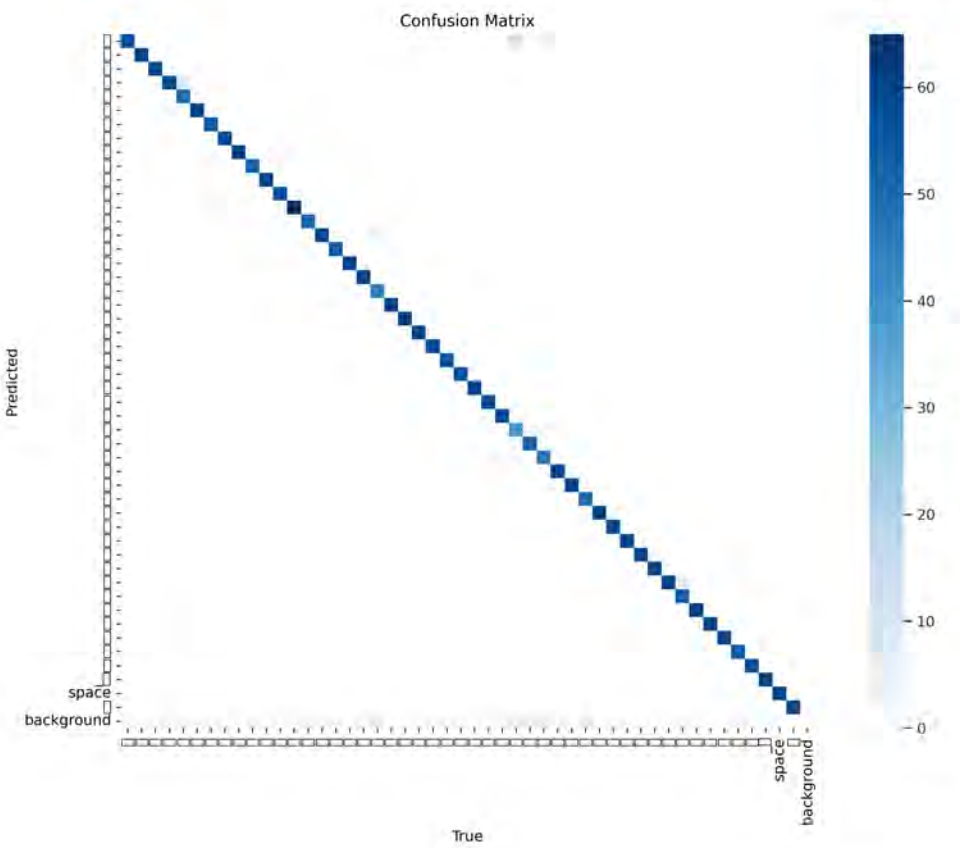


Figure 4.13: Confusion Matrix For Validation

Here we can see the validation precision curve

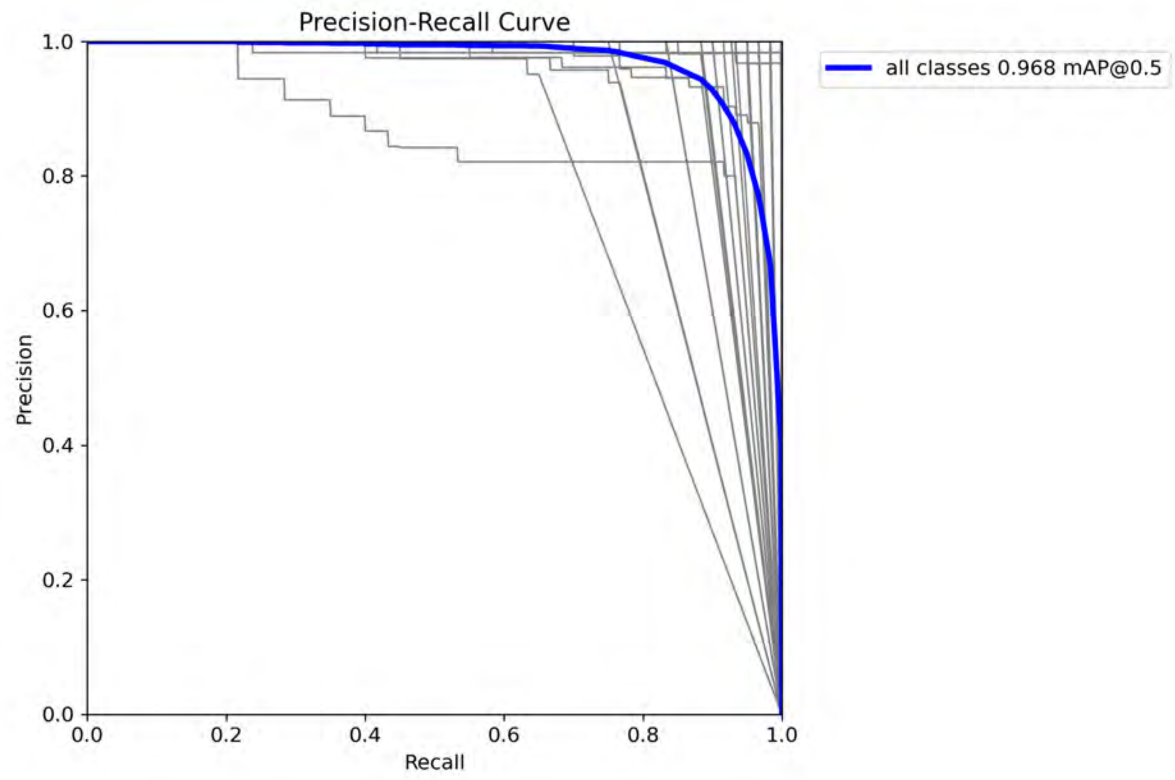


Figure 4.14: Validation Precision-Recall Curve

Here is the validation F1 curve

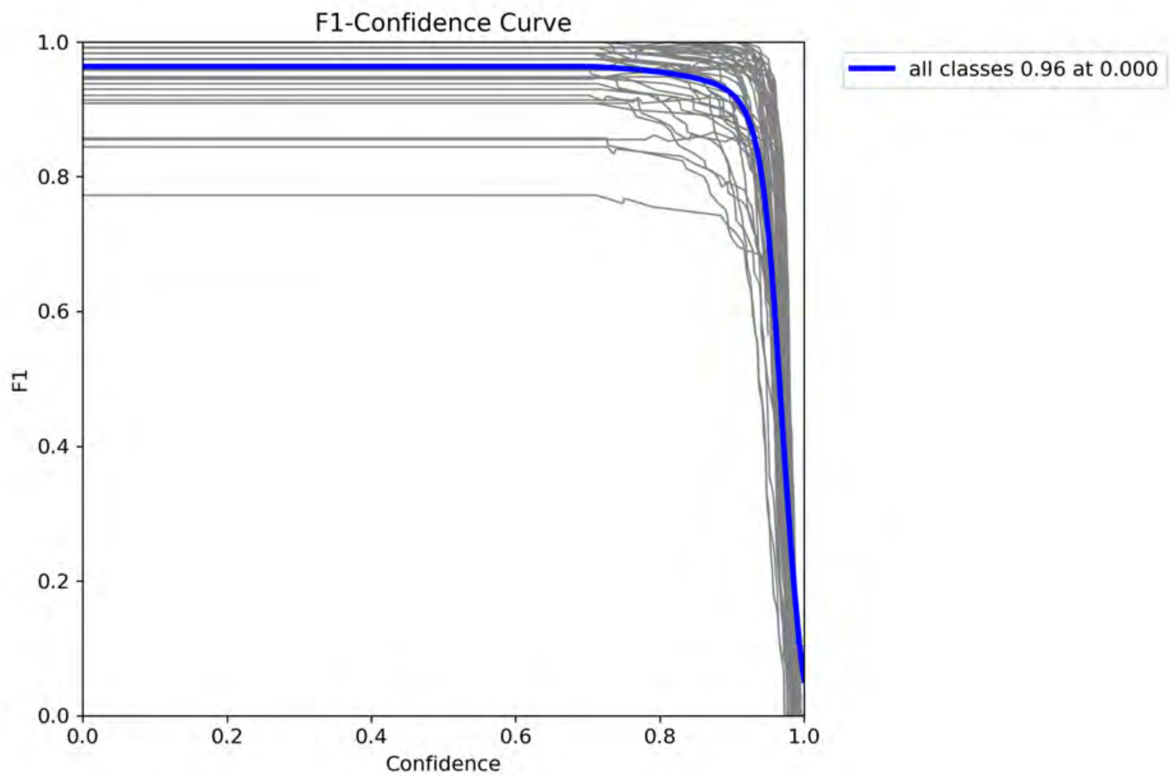


Figure 4.15: Validation F1 Curve

Here we can see the overall result of the training process. It shows about the model being overfitted or underfitted.

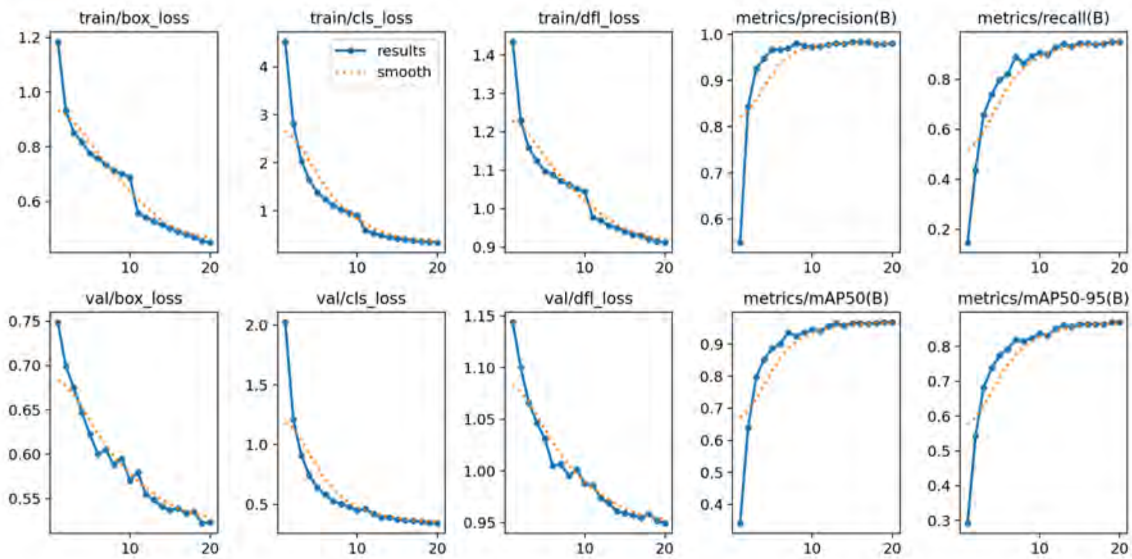


Figure 4.16: Results for overfitting and underfitting

4.2.2 Model Integration

After analyzing and selecting the perfect model for the smart avatar tutor teaching Bangla sign language, we exported the YOLO V8 Nano Baseline model in PyTorch format. Then we converted the model into ONNX model format to integrate with our system. Later, we developed a Python desktop app using opencv-python, onnxruntime-gpu, and PyQt5 to develop the smart avatar tutor system. The smart avatar tutor system initially displays a character sign language image with the character name followed by the accuracy on the right side of the screen. Then the user opens his camera to mimic the Bangla sign language character. Initially, the threshold is 70%. Due to the environment and lighting being different from the dataset we trained, the accuracy does not always give the optimal result. Therefore, there might be some discrepancy in the accuracy. However, our system was able to detect up to 95% in an environment where there is a proper amount of lighting. So, the system is quite user-friendly to use and learn all 49 Bangla Sign Language characters.

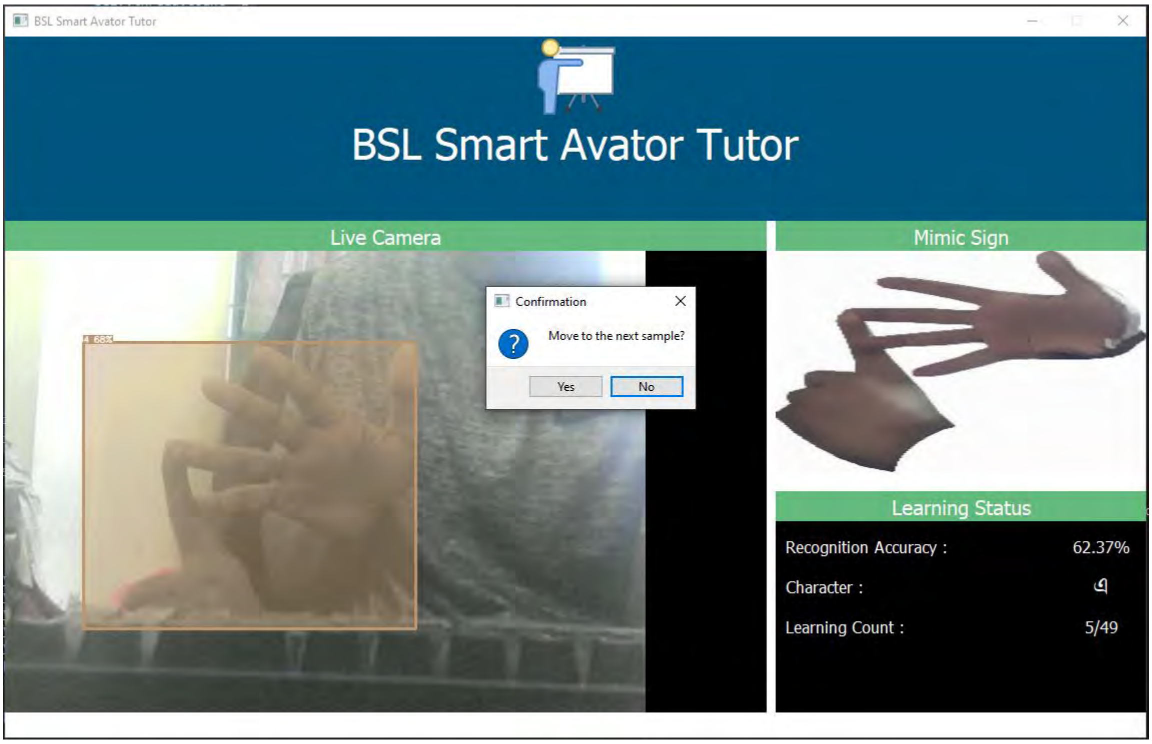


Figure 4.17: Smart Avatar Tutor example 1

.

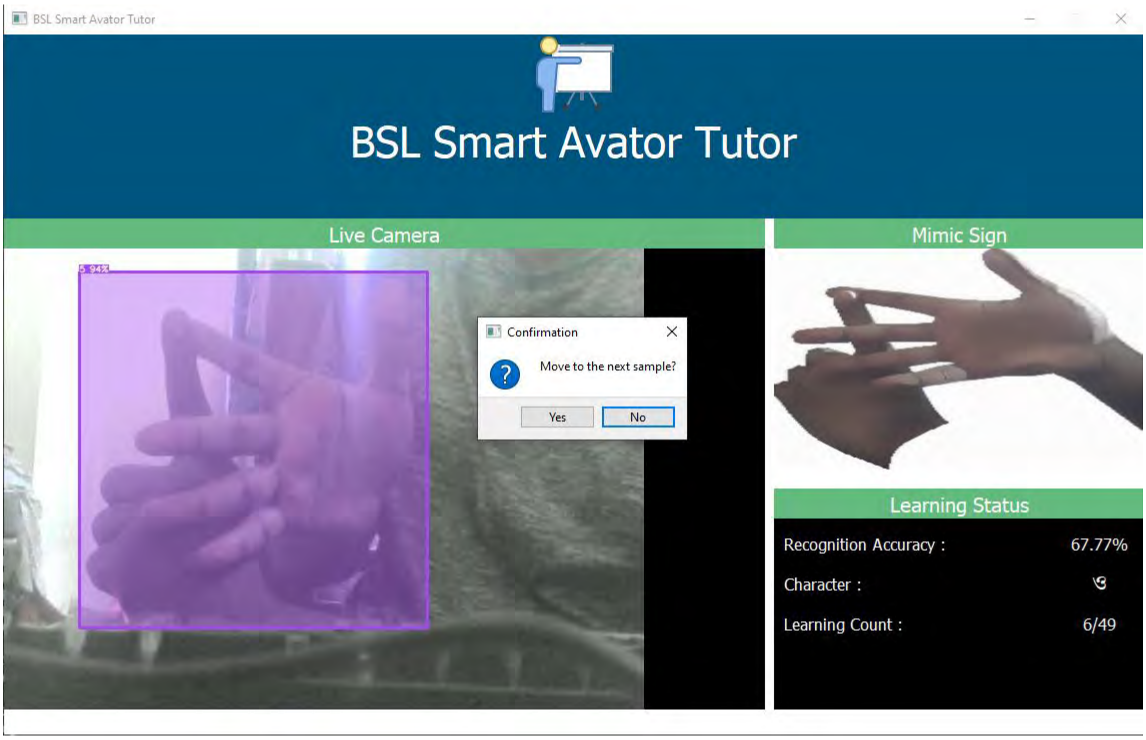


Figure 4.18: Smart Avatar Tutor example 2

. Initially, the user mimics the sign shown by the avatar tutor then once the user reaches a certain threshold the system asks whether he wants to go to the next character. He can choose to stay at the current level and analyze the Bangla Sign Language character even more. Eventually, he can move to the next character and continue his learning. The smart avatar tutor provides effective learning as well as an interactive User Interface that makes it more interesting to the user.

Chapter 5

Conclusion

5.1 Conclusion

Our main aim was to implement a teaching system for the hearing-impaired community in Bangladesh. We have observed that proficiency in Bangla sign language is the only way to bridge this communication gap. We focused on integrating the Yolov8n baseline model with our system for this purpose as based on our research this model outperforms all the other standard models for real-time object detection. The model was trained with 100 epochs on a dataset consisting of 49 signs- alphabets and characters from the Bangla language. The addition of an avatar tutor is a novelty in our research for teaching BSL. Throughout this study, we tried to meet certain objectives including optimizing the avatar for accuracy and efficiency and creating of an immersive learning environment.

In our thesis, we have tried to build a system that will be user-friendly and engaging for the learners. However, there is room for improvement in our work which we can implement in the future. Our system was able to detect 95% of hand gestures with proper lighting. However, the accuracy varied based on the environment, lighting, and the complexity of the hand gesture, which we would like to work on to improve further. In the future, we would like to gather feedback from the users and based on their first impressions, we could modify our system accordingly.

5.2 Future Work

While researching the possibilities of our work, we came across various ways to improve our work to the next level. Our research "A Smart Avatar Tutor for Mimicking Characters of Bangla Sign Language," future work can explore the other possibilities of enhancing the capabilities of the avatar tutor. Some of the key areas that can make a significant contribution to the future work include:

Developing the gesture recognition capabilities, for example, building augmented reality Future efforts could concentrate on refining the gesture recognition capabilities specific to BSL. This might involve developing more advanced algorithms or incorporating new technologies like augmented reality (AR) and Virtual Reality (VR) to make the learning experience more interactive and accurate for BSL gestures. Furthermore, expanding BSL Content and Curriculum and Broadening the field of Bangla sign language with word recognition and sentence recognition. This advancement will add more benefits for the learners. In addition, making the learning strategy dynamic for the user can be a great improvement to this work. This will enable users to schedule or organize their lessons in a manner that can help them enhance their learning in their way. This will make Bangla Sign Language learning more effective. Developing dedicated mobile and web applications for the avatar tutor focused on BSL can increase accessibility and convenience, allowing learners to engage with the tool across different devices and settings. By focusing on these specific areas, future work can significantly enhance the avatar tutor's effectiveness and reach, making it a more comprehensive and impactful tool for BSL education in Bangladesh.

Bibliography

[1] C. Savur and F. Sahin, "American sign language recognition system by using surface emg signal," Oct. 2016, pp. 002 872–002 877. DOI: 10.1109/SMC.2016.7844675.

[2] M. Rahaman, M. Jasim, M. Ali, and M. Hasanuzzaman, "Bangla language modeling algorithm for automatic recognition of hand-sign-spelled bangla sign language," *Frontiers of Computer Science (electronic)*, vol. 14, Aug. 2018. DOI: 10.1007/s11704-018-7253-3.

[3] M. Islam, M. M. Rahman, M. H. Rahman, M. Arifuzzaman, R. Sassi, and M. Aktaruzzaman, "Recognition bangla sign language using convolutional neural network," Sep. 2019, pp. 1–6. DOI: 10.1109/3ICT.2019.8910301.

[4] M. S. Islam, S. Sultana Sharmin, A. S. A. Rabby, and S. Hossain, "A simple and mighty arrowhead detection technique of bangla sign language characters with cnn," in Jul. 2019, pp. 429–437, ISBN: 978-981-13-9180-4. DOI: 10.1007/978-981-13-9181-1_38.

[5] L. Quandt, J. Lamberton, A. Willis, *et al.*, "Teaching asl signs using signing avatars and immersive learning in virtual reality," Oct. 2020. DOI: 10.1145/3373625.3418042.

[6] L. M. Haile, A. U. Orji, K. Kamenov, P. S. Briant, J. D. Steinmetz, A. Abdoli, *et al.*, "Hearing loss prevalence and years lived with disability, 1990–2019: Findings from the global burden of disease study 2019," *The Lancet*, vol. 397, no. 10278, pp. 996–1009, 2021. DOI: 10.1016/S0140-6736(21)00516-X. [Online]. Available: https://dx.doi.org/10.1016/S0140-6736(21)00516-X.

[7] C. Lee, K. K. Ng, C.-H. Chen, H. Lau, S. Chung, and T. Tsoi, "American sign language recognition and training method with recurrent neural network," *Expert Systems with Applications*, vol. 167, p. 114 403, Apr. 2021. DOI: 10.1016/j.eswa.2020.114403.

[8] S. Patil, Y. Shah, P. Narkhede, A. Thakare, and R. Pitale, "Gesture detection using tensor flow lite efficient net model for communication and e-learning module for mute and deaf," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 10, no. 8, 2021. DOI: 10.35940/ijitee.h9204.0610821. [Online]. Available: https://dx.doi.org/10.35940/ijitee.h9204.0610821.

[9] S. M. Rayeed, G. Akram, S. Tuba, G. Zilani, H. Mahmud, and M. K. Hasan, "Bangla sign digits recognition using depth information," Mar. 2022, p. 36. DOI: 10.1117/12.2623400.

[10] X. Jia, Y. Wang, and T. Chen, "Forest fire detection and recognition using yolov8 algorithms from uavs images," in *2023 IEEE 5th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, 2023, pp. 646–651. DOI: 10.1109/ICPICS58376.2023.10235675.

[11] S. S. Khan, A. Hasib, J. F. Eva, *et al.*, *Bdsl 49: A comprehensive dataset of bengali sign language*, Apr. 2023. [Online]. Available: https://data.mendeley.com/datasets/k5yk4j8z8s/6.

[12] D. Kothadiya, C. Bhatt, A. Rehman, F. Alamri, and T. Saba, "Signexplainer: An explainable ai-enabled framework for sign language recognition with ensemble learning," *IEEE Access*, vol. PP, pp. 1–1, Jan. 2023. DOI: 10.1109/ACCESS.2023.3274851.

[13] G. Ministry of Social Welfare. "Disability information system." Published on 21 May 2023. (2023), [Online]. Available: https://www.dis.gov.bd (visited on 05/24/2023).

[14] N. Nath, R. Md, Z. Dhruba, and K. Elahi, "Bangla hand sign recognition using convolutional neural network and transfer learning," Jan. 2023.

[15] F. N. Ortataş and M. Kaya, "Performance evaluation of yolov5, yolov7, and yolov8 models in traffic sign detection," in *2023 8th International Conference on Computer Science and Engineering (UBMK)*, 2023, pp. 151–156. DOI: 10.1109/UBMK59864.2023.10286611.

[16] J. Terven and D.-M. Cordova-Esparza, "A comprehensive review of yolo: From yolov1 to yolov8 and beyond," Apr. 2023.

[17] S. Das, M. S. Imtiaz, N. H. 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*, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417422019327 (visited on 05/24/2023).