# ENHANCING ASSISTIVE SIGN LANGUAGE CONVERTER FOR INDIVIDUALS WITH HEARING AND SPEECH IMPAIRMENTS

UNDERGRADUATE RESEARCH THESIS SUBMITTED
IN PARTIAL FULFILMENT OF THE REQUIREMENTS OF THE DEGREE OF
BACHELOR OF THE SCIENCE OF ENGINEERING

**Submitted by:**
GUNARATHNA L.P.N. (2020/E/046)
SOMARATHNA S.V.A.P.K. (2020/E/212)

**DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING**
**FACULTY OF ENGINEERING**
**UNIVERSITY OF JAFFNA**
**OCTOBER 2025**

# ENHANCING ASSISTIVE SIGN LANGUAGE CONVERTER FOR INDIVIDUALS WITH HEARING AND SPEECH IMPAIRMENTS

**Supervisor(s):**

Supervisor Name1 : Dr.T.Mukunthan (Department of Electrical and Electronic Engineering, University of Jaffna)

Supervisor Name2 : Prof. M. K. Ahilan (Department of Electrical and Electronic Engineering, University of Jaffna)

Supervisor Name3 : Mr.R. Valluvan (Department of Electrical and Electronic Engineering, University of Jaffna)

**Examination Committee:**

Lecturer 1 ..........................................

Lecturer 2 ..........................................

# KEYWORDS

Sign Language Translation (SLT)

Sign Language Recognition (SLR)

Sinhala Sign Language (SSL)

How2Sign Dataset

Deep Learning

Neural Networks

Transformer Model

Long Short-Term Memory (LSTM)

Gated Recurrent Unit (GRU)

Hybrid LSTM-Transformer Model

Feature Extraction

I3D Features (Inflated 3D ConvNet)

Mediapipe Skeleton Extraction

Temporal Modeling

Sequence-to-Sequence Learning

Natural Language Processing (NLP)

Encoder-Decoder Architecture

# ABSTRACT

The increasing demand for inclusive communication technologies has driven research into systems that can bridge the gap between individuals with hearing and speech impairments and the general population. This research presents the development of an assistive sign language translator that converts sign gestures into meaningful text and speech using deep learning and computer vision techniques. The system aims to enhance accessibility in healthcare and public service environments by enabling real-time communication between hearing-impaired individuals and others.

The solution utilizes the How2Sign dataset and applies pose estimation using i3D for feature extraction, followed by a Transformer-based model for sentence-level translation. Comparative experiments were conducted with LSTM, GRU, and hybrid Transformer architectures, evaluating their performance based on accuracy and BLEU score metrics. The results demonstrated that the Transformer model achieved superior translation accuracy and fluency, outperforming recurrent models in both training and real-time testing scenarios. The proposed Transformer model obtained a BLEU score of 0.0478, indicating its effectiveness in generating more accurate and coherent sign-to-text translations during both training and real-time testing.

The implementation further includes a prototype mobile application that provides user-friendly interaction, enabling real-time translation from video input to text and speech output. This project significantly contributes to promoting social inclusion and accessibility, particularly within medical and educational contexts. The findings highlight the potential of technological innovations to enhance communication between hearing and speech-impaired individuals and others, supporting the wider adoption of assistive solutions in Sri Lanka and beyond.

# DECLARATION

We, the undersigned, hereby declare that this report was written by ourselves and the work contained therein is our own, except where explicitly stated in the text.

Student Name 1 : GUNARATHNA L.P.N. (2020/E/046)
Student Name 2:  SOMARATHNA S.V.A.P.K. (2020/E/212)

## CONTRIBUTION TO THE FINAL THESIS BY THE MEMBERS IN GROUP

| Sections | 2020/E/046 | 2020/E/212 |
|---|---|---|
| **CHAPTER 1: INTRODUCTION** | | |
| 1.1 Motivation and Overview | 🟦 | 🟦 |
| 1.2 Aims and Objectives | | 🟦 |
| 1.3 Thesis Scope | 🟦 | 🟦 |
| **CHAPTER 2: LITERATURE REWIEW** | | |
| 2.1 Introduction | 🟦 | |
| 2.2 Forecasting Models | | 🟦 |
| 2.2.1 Computer vision based approaches | | 🟦 |
| 2.2.2 IOT based approaches | | 🟦 |
| 2.2.3 Hybrid approaches | | 🟦 |
| 2.3 Performance Analysis | 🟦 | |
| 2.2.1 Computer vision based approaches | 🟦 | |
| 2.2.1 IOT based approaches | 🟦 | |
| 2.2.1 Hybrid approaches | 🟦 | |
| **CHAPTER 3 : METHODOLOGY** | | |
| 3.1 Methodology | 🟦 | |
| 3.2 Detailed Methodology | 🟦 | |
| 3.2.1 Study of Sign Language Basics and environment | 🟦 | |
| 3.2.2 Selection of Approach | 🟦 | |
| 3.2.3 Finalize the use case | | 🟦 |
| 3.2.4 Review of Existing Projects | | 🟦 |
| 3.2.5 Technical Preparation | | 🟦 |
| 3.2.6 Data Acquisition | | 🟦 |
| 3.2.7 Data Preprocessing And Augmentation | 🟦 | |
| 3.2.8 Feature Extraction | 🟦 | |

| | | |
|---|---|---|
| 3.2.9 Model Development | 🟦 | 🟦 |
| 3.2.10 Model Evaluation and Improvement | | 🟦 |
| 3.2.11 Mobile App Deployment | 🟦 | |
| **CHAPTER 4: EXPERIMENTAL FRAMEWORK** | | |
| 4.1 Introduction | | 🟦 |
| 4.2 Experimental Framework | | 🟦 |
| 4.2.1 Initial Experiments | | 🟦 |
| 4.2.2 Transition to Sentence-Level Data | 🟦 | |
| 4.2.3 Model Development | 🟦 | |
| 4.2.4 Real-Time Implementation | 🟦 | |
| **CHAPTER 5: EXPERIMENTAL RESULTS AND ANALYSIS** | | |
| 5.1 Introduction | 🟦 | |
| 5.2 Experimental Result | | 🟦 |
| **CHAPTER 6: CONCLUSIONS AND FUTURE DIRECTIONS** | | |
| 6.1 Summary and Conclusion | 🟦 | |
| 6.2 Future Work | | 🟦 |
| **REFERENCE** | 🟦 | 🟦 |

# TABLE OF CONTENT

KEYWORDS ........................................................................................................2

ABSTRACT........................................................................................................3

DECLARATION ................................................................................................4

CONTRIBUTION TO THE FINAL THESIS BY THE MEMBERS IN GROUP ..................5

TABLE OF CONTENT ........................................................................................7

LIST OF FIGURES ..............................................................................................9

LIST OF TABLES..............................................................................................10

ABBREVIATIONS AND ACRONYMS .................................................................11

ACKNOWLEDGEMENT .....................................................................................13

**Chapter 1: INTRODUCTION ..................................................................14**

1.1Motivation and Overview ................................................................................14

1.2 Aims and Objectives .......................................................................................15

1.3 Thesis Scope ..................................................................................................16

**Chapter 2: LITERATURE REWIEW ...........................................................17**

2.1 Introduction.....................................................................................................17

2.2 Forecasting Models .........................................................................................18

______2.2.1 Computer vision based approaches .....................................................18

______2.2.2 IOT based approaches.........................................................................25

______2.2.3 Hybrid approaches ..............................................................................26

2.3 Performance Analysis ......................................................................................27

______2.3.1 Computer vision based approaches .....................................................27

______2.3.2 IOT based approaches.........................................................................34

______2.3.3 Hybrid approaches ..............................................................................35

**Chapter 3: Methodology..................................................................................37**

3.1 Methodology....................................................................................................37

3.2 Detailed Methodology .....................................................................................37

3.2.1 Study of Sign Language Basics and environment .......................................37

3.2.2 Selection of Approach ..................................................................................38

3.2.3 Finalize the use case ....................................................................................38

3.2.4 Review of Existing Projects...........................................................................39

3.2.5 Technical Preparation ..................................................................................39

3.2.6 Data Acquisition ..........................................................................................40

3.2.7 Data Preprocessing And Augmentation....41
3.2.8 Feature Extraction....45
3.2.9 Model Development....47
3.2.10 Model Evaluation and Improvement....50
3.2.11 Mobile App Deployment....51

**Chapter 4: Experimental Framework ....55**

4.1 Introduction....55

4.2 Experimental Framework....56

4.2.1 Initial Experiments....58

4.2.2 Transition to Sentence-Level Data....60

4.2.3 Model Development....62

4.2.4 Real-Time Implementation....63

**Chapter 5: Experimental Results and Analysis ....65**

5.1 Introduction....65

5.2 Experimental Result....66

**Chapter 6: CONCLUSIONS AND FUTURE DIRECTIONS ....78**

6.1 Summary and Conclusion....78

6.2 Future Work....79

REFERENCES....81

# LIST OF FIGURES

Figure 1 : Optical Flow vector Analysis ..... 43

Figure 2 : Optical Flow vector Calculation ..... 44

Figure 3 : Image Augmentation ..... 44

Figure 4 : Skeletal Representation of Human Gesture Key Points ..... 45

Figure 5 : Stylized Skeletan Figure ..... 45

Figure 6 : Overall Architecture of I3D ..... 47

Figure 7 : Overall Architecture of GRU model ..... 48

Figure 8 : Overall Architecture of LSTM model ..... 49

Figure 9 : Overall Architecture of Transformer model ..... 50

Figure 10 : Proposed System ..... 52

Figure 11 : Proposed Architecture ..... 53

Figure 12 : Proposed Mobile Application UI ..... 54

Figure 13 : Engaging with Hearing and Speech Impaired Students for Data Collection ..... 57

Figure 14 : Patient-Side Sign Language Video Capture and Transmission ..... 63

Figure 15 : Doctor-Side Sign Language Video Processing and Translation ..... 64

Figure 16 : Visualization of model performance through a confusion matrix, comparing predicted labels to actual labels for each word. ..... 67

Figure 17 : Real-Time Sign Language Prediction Using LSTM Model ..... 68

Figure 18 : Training and Validation Accuracy Comparison for LSTM model ..... 69

Figure 19 : Training and Validation Accuracy Comparison for Transformer model ..... 69

Figure 20 : For the same video, a comparison of the original I3D features vs. the custom I3D features extracted by our model ..... 70

Figure 21 : Original Video Frame with Background And Processed Video Frame with Background Removed ..... 73

# LIST OF TABLES

Table 1 : Comparison of GRU, LSTM, and Transformer-Based Models for Sign Language Translation ........ 48

Table 2 : Sample Predictions and BLEU Scores for Encoder + T5 Decoder on the HOW2SIGN Dataset ........ 73

Table 3 : Sample Predictions and BLEU Scores for Encoder + Transformer Layer + Decoder on the HOW2SIGN Dataset ........ 74

Table 4 : Sample Predictions and BLEU Scores for Encoder + LSTM Layer + Decoder on the HOW2SIGN Dataset ........ 76

Table 5 : Comparison of BLEU Scores with State-of-the-Art (SOTA) ........ 77

# ABBREVIATIONS AND ACRONYMS

| | | |
|---|---|---|
| ASL | : | American sign language |
| AUC | : | Accuracy and under the curve |
| BiLSTM | : | Bi-directional Long Short Term Memory |
| BSL | : | British Sign Language |
| CNN | : | Convolutional Neural Network |
| CRF | : | Conditional Random Fields |
| CTC | : | Connectionist Temporal Classification |
| D | : | Deletion |
| Gloss2Text | : | Gloss annotations to text |
| GRU | : | Gated recurrent unit |
| GT | : | Ground truth |
| HCI | : | Human-computer interface |
| HMM | : | Hidden Markov Model |
| I | : | Insertion |
| ISL | : | Indian Sign Language |
| LDA | : | Linear Discriminant Analysis |
| LLM | : | Large language model |
| LR | : | Logistic Regression |
| LSTM | : | Long Short-Term Memory |
| MS | : | Multi-speaker |
| NB | : | Naïve Bayes |
| NIE | : | National Institute of Education |

| | | |
|---|---|---|
| NLP | : | Natural language processing |
| PCA | : | Principal Component Analysis |
| POS | : | Part-of-Speech |
| PTN | : | Pose Transformer Network |
| RC | : | Ridge Classifier |
| RFC | : | Random Forest Classifier |
| ROC | : | Received operating characteristic |
| S | : | Substitution |
| SD | : | Speaker-dependent |
| SLSL | : | Sri Lankan sign language |
| SLT | : | Sign Language Translation |
| SMC | : | spatial multi-cue |
| SOV | : | Subject + Object + Verb |
| SSL | : | Sri Lankan Sign Language |
| STMC | : | Spatial-Temporal Multi-Cue |
| SVD | : | Singular Value Decomposition |
| TMC | : | Temporal multi-cue |
| UDP | : | User Datagram Protocol |
| WER | : | Word Error Rate |

# ACKNOWLEDGEMENT

We would like to express our sincere gratitude to all those who have supported and guided us throughout the completion of this research project.

First and foremost, we extend our deepest appreciation to our primary supervisor, Dr. T. Mukunthan, from the Department of Electrical and Electronic Engineering, University of Jaffna, for his invaluable guidance, encouragement, and continuous support throughout every stage of this work. His expert advice and constructive feedback were instrumental in the successful completion of this research.

We would also like to express our heartfelt thanks to our co-supervisors, Prof. M. K. Ahilan and Mr. R. Valluvan, from the Department of Electrical and Electronic Engineering, University of Jaffna, for their valuable insights, suggestions, and constant encouragement, which greatly contributed to improving the quality of this study.

Our sincere thanks also go to the Department of Electrical and Electronic Engineering, Faculty of Engineering, University of Jaffna, for providing the necessary facilities, resources, and a supportive environment for our research work.

We are also deeply thankful to the teachers and staff of the schools for the blind and deaf located in Jaffna and Badulla for their valuable time, cooperation, and insightful discussions. Their guidance in helping us understand the real-world challenges and communication difficulties faced by deaf students was truly inspiring and significantly contributed to the social relevance of our research.

We are grateful to our colleagues and friends for their kind assistance, motivation, and cooperation during the development and testing phases of this project.

Finally, we would like to express our heartfelt gratitude to our parents and families for their unconditional love, patience, and constant moral support, which have been our greatest strength throughout this journey.

# Chapter 1: INTRODUCTION

## 1.1 Motivation and Overview

The ability to communicate is an important aspect of human life, essential for interaction, education and employment. However, individuals with speech and hearing impairments face many difficulties while they try to communicate with other members of the society at large. In their attempts to talk with other members of the society at large, hearing and speech impaired people face a great deal of difficulties. Traditional methods as notes written on paper or lip-reading depend on the skills of individuals that can be affected by things like voices, environment. On the other hand, speaking messages might be too long-winded and unworkable for real time conversations. Unfortunately, this language cannot assist communication if someone who does not understand the language uses it thus making deaf people feel unsatisfied and isolated.

However, these limitations have been partially addressed by the introduction of sign language translating assistive technologies which could overcome these barriers sometime in future. Several tools presently available have problems concerning accuracy when it comes to reading sign languages, some are unable to interpret signs well resulting in misunderstandings. This requires fast translation but speed is another burning issue here. Many systems fail to keep up with the pace of conversation flow during real-time communication. In addition, usability counts, complicated and difficult-to-use devices or applications may discourage regular use, especially of people who are not good at technology.

These are the limitations that should be factored in when looking at the need to improve assistive sign language translators. By leveraging advancements in deep learning, computer vision and natural language processing (NLP), we can develop tools that give faster and more accurate translations of sign languages. These improvements will significantly improve user experience thus making the tools more sustainable for everyday usage. The ultimate goal is to establish a communication space where speech and hearing-impaired individuals can easily interact with the rest of the society without difficulty.

These technologies, if developed, can have certain benefits. Thus, the means of sign language translation that are more accurate and able to translate signs faster can help to enhance people's communication in different spheres and levels. More user-friendly better designs can make them more accessible for the users and as result – more acquired, can fit the life simplicity. In addressing these needs, our project expects to play a role in helping persons and integrating in to society with hearing and speech impairment in order to have better chance and functional ability at societal equality.

### 1.2 Aims and Objectives

- Enhance smooth interactions between those with hearing impairments and the entire community, increasing inclusivity.
- Create a real-time sign language translator that can convert hand gestures into text in English.
- Design a mobile application that is easy to use and accessible for users.
- Ensure accurate translation and flexibility for dealing with various differences in sign language.
- Overcome technical difficulties including variable lighting and low-resolution input.

The effectiveness of the project includes improving the interactions between the deaf and dumb populations in order to make the society more flexible. So, we will develop an real time sign language translator to change the sign language gestures to English. We will remove communication barriers. One of the more important activities in this strategy is to develop an sensitive mobile app which will be accessible to all users. Several factors have been adopted as follows to enhance the effectiveness of the developed system, as mentioned in the earlier, the primary focus for the system will be to attempt for accurate translations of the content as well as consider different variations of sign language. Furthermore, we will include failures like low resolution scans or images with different backdrops, and the tool should be compatible with such conditions. The ultimate aim is to discover a well-functioning instrument that will provide a lot of light on the day-to-day struggles of the disabled ear and mouth users.

### 1.3 Thesis Scope

The scope of this research focuses on the construction of a sign language translator which assists patients with speech and hearing impairments to communicate with the medical staff. The project is designed to address the unique communication barriers faced by this group of patients during medical consultations and other forms of interactions.

This research will target on how to include and develop strategies of incorporating special medical words on which the patients are expected to communicate with the medical staff like doctors and nurses during their physical visits to the hospital as well as in the telemedicine. Also, the instrument will be compatible with the existing tele systems and the infrastructure of the hospitals, so it can be easily used in these environments.

Because it is concerning the health care, this task will develop and provide the patients with the real time translation services so that proper instructions can be given to them during important medical examinations and other consultations catered at them. The use of real time translation adds on another dimension to the sign language converter, allowing for it to imbue features that are more intuitive for users.

The scope takes more shape by including the measures to be taken against obstacles like lack of sufficient data and modifying the instrument to fit more domains like telehealth and hospitals settings. This will ensure that the instrument is not only useful in enhancing the communication between the patients and the doctors but also adds on its utility value.

# Chapter 2: LITERATURE REWIEW

## 2.1 Introduction

In recent years, Incredible advancements have been made in the development of devices planned for use in speech and hearing-impaired clients. One of the most useful advancements is discovering ways to interpret sign language in order to facilitate communicating with a speech and hearing-impaired persons. Some of these tools prove most useful in the healthcare setting, as accuracy and efficiency are the essential components of delivering excellent service.

This section further explores into new directions of this area of research with respect to the various techniques that have been used to estimate how well these utilities perform, a study of how effectively these tools could sign and interpret, and a quick look at the data sources that they are based on.

There have been attempts that have been made by Researchers in an aim to enhance the quality standards of sign language converters. The following studies will discuss the possibility to evaluate which tools should be used for specific tasks, as well as the kind of data that these tools require and analyse, in order to identify their applicability in practical scenarios and their drawbacks. It focuses on the development of assets required to reduce barriers between the signed and spoken languages that will introduce efficiency across areas of interest such as medicine, schools and the society in general.

That is, the following categorization can be proposed with reference to particular techniques and devices.

- Computer vision-based approaches
- IOT based approaches
- Hybrid based approaches

## 2.2 Forecasting Models

### 2.2.1 Computer vision-based approaches

A web-based app for remote health consultation and Machine learning model for sign classification for COVID-19 was created by Maria Seraphina Astriani and Marcell Alvianto from the Department of Computer Science, School of Computing and Creative Arts, Bina Nusantara University, Indonesia using BISINDO data [1]. Their methodology involves several key steps, requirements, which should be fulfilled in order to have a positive outcome include installation of OpenCV and MediaPipe Holistics for Data Pre-processing and Data Gathering, how data is obtained for this model, this involves the use of LSTM together with multiple Dense layers for training of this model. The final part involves writing the necessary code and testing it on the model, after the training of the model is over, the model is saved and again loaded again into the system using OpenCV and MediaPipe for prediction in real-time and for testing purposes. The systematic detection is providing for the sign languages of the speech and hearing-impaired individual, such a structure is also useful in establishing a more effective approach since of COVID-19 in this study.

Another reviewed literature presents, as mentioned by Kayo Yin and Jesse Read in Sign Language Translation (SLT), Transformers present an upgrade to the translation system [2]. They proposed an SLT workflow that is completely new and utilizes the Spatial-Temporal Multi-Cue (STMC) and Transformer networks. The experiments in their work based on the 'PHOENIX-Weather 2014T' dataset and proved that end to end with translation and the predicted glosses fares better than the one translated using the ground truth (GT) glosses. This of course shows that SLT can still be enhanced even more by either integrating the simultaneous training of SLR and the translation component or improving the gloss labelling of annotation. The STMC network parses the input video based on the spatial features through the spatial multi-cue (SMC) module and in the next step learns about the temporal correlation using the temporal multi-cue (TMC) module. These features are then passed through the Bi-directional Long Short-Term Memory (BiLSTM) and Connectionist Temporal Classification (CTC) components for sequence learning and decodification. For translation, they employed two stairs Transformer model. Some of the experiments they undertook consisted of mapping GT gloss annotations to text (Gloss2Text) and a full-scale Sign-to-Sign translation employing the STMC-Transformer. (Sign2Gloss2Text).

To improve the response of SLT, a new model, namely Spotter+GPT, was proposed by Ozge Mercanoglu Sincan, Necati Cihan Camgoz, and Richard Bowden [3]. Their method involves a two-step process, first, in the case of dedicated work employing a sign spotter based on a linguistic sign language dataset ( MeineDGS) for sign segmentation of continuous sign video data. The spotter, for sequences of repetitive frames, utilizes sliding window with a frame size of 8, based on the I3D model. As soon as glosses are found, they are handed to a pretrained large language model (LLM), which is chosen to be a ChatGPT model in the current work and these glosses are converted to coherent spoken language sentences. This approach enhances SLT, especially by identifying sign language gestures in real-time and converting them into speech that can be easily processed by individuals with hearing and speech impairments based on the features of computer vision and natural language processing.

In the work done recently by Neena Aloysius and M. Geetha different forecasting models were used in solving different issues to do with continuous sign language recognition (CSLR) [4]. The authors paid attention to the new approaches which employed the data from the vision system rather than the dependence on sensors. They looked at approaches like Hidden Markov Models (HMM), Conditional Random Fields (CRF) and Dynamic Time Warping (DTW). These traditional models establishing from isolated sign language recognition (ISLR) experienced reduced performance in CSLR because of the challenge in locating Movement Epenthesis (ME) segments amidst the flow of other segments. As of the Development of Deep Learning technologies, there are efficient solutions for vision-based CSLR. An overview of these methods was provided in this study together with the assessment of their efficiency and the identification of a list of predispositions that carry out the basic assumptions of vision-based CSLR systems.

Sampada Wazalwar and Urmila Shrawankar has created a system to overcome the problem of different levels of communication by incorporating various modules to include major features of sign language translation where a database consisting of Indian Sign Language of medical words and phrases has been implemented [5]. For sign recognition, the Vision Processing Algorithm used is a Hidden Markov Model (HMM) coupled with a Haar cascade classifier. This makes it possible for the system to recognize signs and further interpret them as appropriately needed. The interpretation module employs language technology necessarily the Part-of-Speech (POS) tagging and parsing for being able to have better understanding and context while translating sign language into spoken or written language. This discussion reveals that their strategy unites the computer vision and natural language processing techniques to help sufferers of hearing and speech impairment in communication, especially in the clinical setting.

Researchers, Y. Zhao, X. Zhang, R-S. Hu, J. Xue, X. Li, L. Che, R. Hu, and L. Schopp from the Department of Computer Science, University of Missouri, Columbia, have proposed and implemented an automatic closed captioning system to facilitate telemedicine to the hearing and speech impaired [6]. It is this, Their system uses the newest large vocabulary continuous speech recognition software. Some of the sub-modules are speech stream separation, acoustic model, N-gram language model, real-time decoding, confidence annotation, and human-computer interface (HCI). As for data collection they relied on telemedicine network of University of Missouri where they got record of 51 hours where 7 healthcare providers participated. It captures the audio or video stream of the conversation between the doctor and the patients, extracts the doctor’s speech, and feeds the extracted speech to the automatic speech recognition system which then displays the recognized words along with their confidence levels. The acoustic model entails a form of a Gaussian mixture density Hidden Markov Model (HMM). Its language model is trained from limited transcriptions of telehealth conversations and complemented by such data from other domains. The conversational style of the medical consultation is impulse. This approach ensures that closed captioning is as right and promptly given as possible, thus improving telemedicine usability by the hearing-speech impaired persons.

Some of the researchers who have implemented a perfect solution are Mathieu De Coster, Mieke Van Herreweghe, and Joni Dambre, they have used feature extraction via OpenPose, for human key point estimation, and end-to-end feature learning with Convolutional Neural Networks (CNNs) [7]. They used the multi head attention technique originating from transformers to detect single signs from the Flemish Sign Language corpus. The proposed method has better pass rate result compared to previous state-of-the-art techniques.

They explored four methods for isolated sign language recognition (SLR) such as PoseLSTM, Pose Transformer Network (PTN), Video Transformer Network, and Multimodal Transformer Network. PoseLSTM has been proposed for action understanding and it is based on Long Short-Term Memory (LSTM) networks as well as features from OpenPose by default. Classification of multiple head attention classifier devoid of OpenPose key points in PTN. The Video Transformer Network uses a framewise 2D CNN that extracts from each frame a 512- dimensional vector which then goes through the multi-head self-attention network. The classification in the Multimodal Transformer Network is done with OpenPose key points while the learned features are combined and subjected to a feature extractor of the pre-trained ImageNet.

All methods were trained and evaluated on the same data set which was further divided on training (70%), validation (10%), and the test (20%) set, having total images 13777, 7433, and 3910 respectively out of which 10777 images were used for training, 743 images for validation, and 1910 images for testing. The models were trained by applying random temporal cropping thus providing data augmentation for temporal data. Both the training and the evaluation were performed using PyTorch 1.3. on Nvidia GeForce GTX 1080 Ti GPUs whereby each GPUs. It has been shown to enhance isolated sign language recognition accuracies, and this has been demonstrated by the significant improvements.

Authors from the University of Mysore, B. M. Chethana Kumara and H. S. Nagendraswamy, have devoted their recent study to the improvement of sign language recognition with the help of new approaches [8]. They used the Local Binary Pattern

Variance (LBPV) for texture description which is a descriptive method that lacks quantization in grayscale images capturing the local contrast therein. This method entailed converting color frames of sign videos to the grayscale domain and using LBPV to obtain fine texture features. Variability due to different signers and conditions needed to be addressed, for which they used K-means clustering to select frames out of each video to obtain more consistent features. Besides, they used hierarchical clustering for the control of intra-class dispersion, which means that several similar signs were grouped into one class. Their approach enabled accurate sign recognition to the relations between a test sign's features and those in a knowledgebase leveraging symbolic similarity assessments. This research is a contributed towards the advancement of technology in the area of sign language recognition with many variations in signing and different environments.

Other recent work is by Amit Moryossef of Google and colleagues, studying pose estimation models for sign language recognition (SLR) with researchers from the University of Surrey and University of Zurich [9]. They preferred skeletal patterns extracted from OpenPose and MediaPipe Holistic models, which provided low-dimensional and privacy-preserving representations of human movements. These representations abstract over appearance and background to afford a measure of invariance well suited to motion recognition tasks like SLR. The teams participated in the CVPR21 ChaLearn challenge using these pose estimation outputs to develop three systems: The first proposed approach is based on the OpenPose approach, the second one is based on the Holistic approach, and the last one integrates all the two approaches. Every team used specialized architectures like SLR transformers and sequence classification models alone with training method, CTC loss and so on. They added that their findings were expected to explore the practicality and feasibility of existing pose estimation technologies in the context of sign language recognition while identifying the opportunities and challenges for the further research of the topic.

Sang-Ki Ko, Heeyeong Jo and Jeonghee Choi, the researchers of Korea Electronics Technology Institute (KETI) introduced the method of using neural network that is capable of recognizing the sign language videos and converting these signs into natural language sentences with the help of large-scale sign language set of KETI, which includes the set of 14,672 high-resolution videos of 524 sign and 14 deaf signers [10]. With this regard, they employed OpenPose to gather 124 key points that include face, hands, and the upper body of the signers where they scaled the outcomes based on variance, background, and differences of the signers. This is their model based on the encoder-decoder neural network that contains an attention mechanism and RNN in the decoder like LSTM or GRU that is used for translation of the sequential data. Moreover, feature vector normalization and random frame skip sampling to the data augmentation process were carried out to improve on the model. Most importantly, this strategy validates the advantages of high-level feature extraction and even further normalization methods in the development of the sign language translation platform.

With a contribution from Laia Tarres, Gerard I. Gallego, Amanda Duarte, Jordi Torres, Xavier Giro-i-Nieto and other researchers from Universitat Politecnica de Catalunya, Barcelona Supercomputing Center and Amazon a lot of progress has been achieved concerning How2Sign, a large-scale dataset in sign language [11]. They trained a Transformer model on I3D video features and adopted the lower BLEU score as the contribution measure. Regarding data preprocessing, they used I3D features extracted from RGB video frames where they pay attention to both visual and temporal characteristics of the frames. For textual manipulation, they made text lower case, tokenized it through the Sentence Piece, and then done post-processing.

Their set up involved a standard Transformer encoder-decoder, theoretically with 6 encoder layers and 3 decoder layers. The encoder recorded the video features of the videos while the decoder worked on the textual descriptions. They set a subsampling of vocabulary to 7000 sub-words, a batch size to 32, and used cross-entropy as the loss function with label smoothing. Considering the facts of several aspects of the training process, which has been described earlier, the following developments were offered: the use of the Adam optimizer with a appropriate warm up learning rate that has a cosine decay. Their training process has included 108 epochs and the time to complete this was about 3.5

hours on a single NVIDIA GeForce RTX 2080 Ti GPU. This work adopted as the new starting reference for translating sign language using the How2Sign dataset.

To investigate if the discourse level influences the translation of sign language, the following researchers from Google collaborated with Rochester Institute of Technology: Garrett Tanzer, Maximus Shengelia, Ken Harrenstien, David Uthus [12]. They offered a new human reference for ASL to English translation as a result of How2Sign dataset , however, they looked at the long-range syntactic relation, thus, unlike the previous work, not concerned with the sentence clipping. Their approach was to employ interpreters who are low level in ASL-English to translate ASL clips without concerning the rest of the document and concluded finding that information that is 33% of the sentences is needed for further clarification in such a way. As a source, the How2Sign test set was selected as they randomly chose 184 ASL translations from 149 How2Sign narratives and mentioned that Deaf annotators translated and rated the accompanied clips and inform that facial expressions and mouthing are crucial in the sign languages. This work highlighted the importance of conversational level context and contributed a number of valuable insights into the discussion of the matters pertaining to sign languages' translation with machine translation.

Patricia Cabot Alvarez Xavier Giro Nieto and Laia Tarres Benet redoing from the recent literature on sign language translation aimed at the How2Sign with a method that is based on the Transformer [13]. First of all, they test it on reproducing the results on the PHOENIX2014T dataset in which the sign features were extracted with the help of a pre-trained CNN. They then altered the process to How2Sign, because they could not find any gloss annotations, so they associated English sentences with frames of the video. Their work was focused on the variety and the relative size of How2Sign compared to the small PHOENIX2014T. They extracted features from each of the sign language videos using an I3D neural network trained for sign language video searching. The findings of this adaptation would prove useful in providing foundation for subsequent researches in sign language translation.

Jana Košecká, Antonios Anastasopoulos and Pooya Fayyazsanavi also contributed to the Gloss2Text translation stage with the help of LLMs, data augmentation methods and label-smoothing loss that better regulates the translation of glosses [14]. They concentrated on transforming gloss annotations into spoken word sequences whereby they adapted and trained fine-tuning LLMs such as NLLB-200 and MT5. They also combined paraphrasing and back translation in order to expand on the data in order to build more reliable models. They adapted this approach to PHOENIX Weather 2014T dataset and proved that their model is more efficient and more than the state of art model in Gloss2Text translation.

### 2.2.2 IOT based approaches

UNS student's group have designed a novel sign language translation glove dubbed "Glova" which is in aid for people with speech and hearing difficulties [15]. In the Glova project, flex sensors on the fingers help the system establish hand movements of letters in representing words. In order to increase the degree of precision, the system uses the Levenshtein distance algorithm for the automatic correction of the text, so that the translated word is a precise match. This technology entails an extensive record of the Indonesian language words and erasures to provide translations and correction simultaneously. This is achieved by using an LCD screen, and a Blynk application to display the translated words or characters and, the errors if any are displayed in another line on the screen. It is therefore an important innovation that helps in clearing the communication barrier to those with challenges of speech and or hearing impairment lacking understanding of sign language.

Introducing the world to American Sign Language and English speech for disabled people was a wearable glove designed by UCLA scientists and engineers [16]. It has stretchable sensors on the fingers that help detect the movements of the hand and the position of the fingers. These movements are translated to electrical signals which are sent to an electrical circuit board fixed on the wrist which in turn sends these movements to an application in a smartphone and the movements are converted to spoken words. There are also adhesives sensors used on the face to point out the facial expressions that make the translation better. The intended concept of this innovation is to enhance communication between signers and non-signers.

### 2.2.3 Hybrid approaches

A group of students participating in the B. Sc in Computer Science program at the University of Colombo has created a Gesture recognition system using Two Myo armbands for recognizing and interpreting sign language found in Sri Lanka [17]. This system is based on signal processing and supervised learning with the use of a restricted manual vocabulary of 49 signs together with three hundred and forty-six different sentences constructed in the SOV (Subject + Object + Verb) pattern. The Myo armbands supplied us with the EMG and IMU data from various sensors, which has been collected and then post-processed and segmented by the extracting each of the signs within the context of a sentence. The primary feature engineering processes that are vital before training a machine learning model are feature selection and feature reduction. This system is to be established to enable sign transmitting since many and simple signs are normally used in day to-day life.

R. M. Rishan, S. Jayalal, T. K. Wijayasiriwardhane from the Faculty of Science University of Kelaniya presenting they designed sensor-based system using Leap Motion, geometry template matching and Natural Language Processing (NLP) for translation from Sri Lankan Sign Language (SSL) to Sinhala text [18]. It consists of a Sign Training Model and a Sign Identifying Model, performed using Leap Motion API and Leap Trainer framework, used to capture and recognise signs. The NLP unit which processes combined signs and signs with multiple meanings used WordNet API for JavaScript as well as the regular expressions. Two data sets which were derived from the help of an SSL interpreter including combined signs and multiple meanings signs using Leap Motion controller were trained. This strategy enhances the probability of making correct form translations to provide coherent meaning of SSL gestures into Sinhala sentences.

Eak-Wei Chong and Boon-Giin Lee of the Keimyung University, Department of Electronics Engineering carried out a study on a sign language recognition prototype using the Leap Motion Controller (LMC) [19]. Unlike many studies that only go for partial SRC from a certain sign language, their study was for a comprehensive one from ASL which has 26 alphabets and 10 digits. They gathered data from 12 university students, the delegates made ASL gestures and simultaneously the LMC captured their hand and finger movements. Simple features including the hand palm sphere radius, position of the hand palm, and position of fingertips was extracted. They used Support Vector Machines (SVM) and Deep Neural Networks (DNN) classifiers to recognize the ASL gestures and compared their performance.

Jordan J. Bird, Anikó Ekárt, and Diego R. Faria provided an analysis on the effect of multimodal approaches in the recognition of ASL signs, where a late fusion was used to show that a model using image classification or Leap Motion data classification outperformed when combined as against individually [20]. They collected a dataset of 18 British Sign Language (BSL) gestures from multiple subjects and benchmarked two deep neural networks: a Convolutional Neural Network (CNN) for vision and an optimized Artificial Neural Network (ANN) for hand Gesture data with the help of Leap Motion. These models were then combined such that they started working in parallel. The late fusion multimodality approach gave better results than single sensor approaches, and it was confirmed that the method of employing transfer learning for the classification of the American Sign Language (ASL). The multimodal model works well and gives the best results when image and 3D hand feature data are used to recognize the sign language.

## 2.3 Performance Analysis

### 2.3.1 Computer vision-based approaches

Maria Seraphina Astriani and Marcell Alvianto from Bina Nusantara University's Computer Science Department have developed a model to classify 11 sign languages, achieving impressive accuracy rates [1]. The performance for training data set, testing data set, and real time classification for the given dataset set as 99%, 98%, and 90%, respectively. The high performance established that the sign language using the LSTM (Long Short-Term Memory) architecture distinguishes the signs and even terms referring to COVID-19. They also used to track and experiment with training and validation accuracy and their loss curves. First the type of activation function and number of epochs were chosen, the required precision was adjusted and the appropriate code was written. Real-time Sign Language recognition is another field where this model can be practically applied which has been clearly depicted by the discussion.

Another research project was improved Sign Language Translation (SLT) system where have provided higher results than the previous techniques by innovating under the supervision of Kayo Yin and Jesse Read [2]. The performance of Their methodology also outperforms possible current benchmarks by more than 5 & 7 BLEU points respectively on the Ground truth along with the predicted glosses of the PHOENIX-Weather 2014T. They also highlighted a amazing improvement of more than 16 BLEU points on the ASLG-PC12 body of work. They also designed it to use transfer learning with English fast Text vectors, which was used to further enhance the performance and also to stabilize the decoder through weight-tying. In addition to that, they propose a novel Transformer Ensemble model which has been reported to perform better than prior works, with a 5 BLEU-4 enhancement on PHOENIX-Weather 2014T and, a brisk 17 BLEU-4 improvement on ASLG-PC12. This piece of work highlights the significance and efficiency of advanced architectures in improving the efficiency and accuracy of sign language translation.

Sign Language Translation (SLT) approach has been invented by Ozge Mercanoglu Sincan, Necati Cihan Camgoz, and Richard Bowden, and the evaluation with BLEU and BLEURT metrics has been made [3]. They used sacreBLEU for BLEU score estimation, BLEURT check points. Further, they used the “Sign Spotting” method on the MeineDGS dataset where they produced decent accuracies on instance and class levels. By fine-tuning the I3D model initially on the BOBSL dataset and subsequently on MeineDGS, they phenomenally reported. The conclusion when applying sign language data in model initialization is that using sign language data in model initialization makes it possible to gain a 1.5% performance increase. They assessed their SLT approach as applied on MeineDGS-V and DGS-20 videos, and improving the Spotter’s performance by tuning the probability thresholds during implementation, which they set at 0. 7 proving the best outcomes with a higher gloss prediction value. Unexpectedly, their system outperformed others specifically on DGS-20 videos, which might be attributed to the fact that the ‘‘ground truth’’ in these videos has a considerably simpler vocabulary structure, which in turn allowed for more evenly pronounced generation of sentence patterns by GPT models. This work shows their creative application of complex knowledge engineering methods in enhancing the realism of translation and recognition platforms.

In their most recent piece of work, Neena Aloysius and M. Geetha did an Excellent performance comparison on the continuous sign language recognition systems [4]. They used the Word Error Rate (WER) which reflect the number of insert from recognitions, deletions, and substitution to convert the hypothesis recognized to the reference sequence. They observed that while previous studies employed Models such as Hidden Markov Models (HMMs) and CRFs, achieved high detection rates in worked up sets but had issues detecting benchmark actual sets. On the downsides, the deep learning models, no matter how they were trained on big data, provided a lower detection rate because of the real-world data complexity and variability. In their work, they stressed the fact that current methods require better models that will be capable of addressing the seemingly complicated issues that are present in sign language capture.

This proves the fact that we must need realistic database to benchmark sign language recognition systems and at the same time it highlights the drawbacks of using conventional techniques rather than deep learning techniques.

Sampada Wazalwar and Urmila Shrawankar have also created a system to help physicians remotely consult with the Deaf since the use of sign language to represent the spoken word so as to be understood by the Deaf is helpful especially in the pandemic period [5]. Its emphasis is on translating individual meaning of the selected keywords of sign language into meaningful phrases. This way the final goal of the patients being treated would be made understood by the physicians in stark clarity and that too without any misinterpretation. This is especially important in healthcare settings. By referring their study, the input keywords from the patient gesture are converted into the output that patients can understand. it improves the reliability of communication, which will help to improve the quality of healthcare away from the premises. In this work, therefore, consumed has redefined their practical use of the technology when used to enhance access and communication during consultations involving deaf people which will enhance reliability in health settings.

Medical speech captioning system has been developed by researchers at the Department of Computer Science at the University of Missouri, Columbia and the developed system have a medical word reference of 46,489 and encounter time of 30.7% [6]. To train and test their system, they used data from five physicians, two females and three males, all whose speech recordings were processed. Hence, five speaker-dependent (SD) and one multi-speaker (MS) acoustic models were trained, with the numerator's five doctors tested against a single MS model. Likewise, one six individual language model and one universal language model for those five doctors were trained. These findings indicate that the captioning word accuracy can be quite different among doctors, at the same time, the desired performance is achieved only within 20% accuracy, and the best option is the ICOW model for all the speakers. These results revealed that the SD models of the different individuals are superior to the MS model, again uniforming the idea of having a normal model for everyone.

Mathieu De Coster, Mieke Van Herreweghe, and Joni Dambre have used ASL-PT as the dataset for evaluating the performance and the results derived from the models and compared it with other techniques such as PoseLSTM and PTN [7]. They noticed a big difference in accuracy, which implies that features derived from OpenPose alone are not sufficiently powerful. This means that improved features can be learned from raw data without following the conventional ways of approaching the problem through a mathematical formula. They assume the motion blur in the available data degrades OpenPose key points measurement and that useful information which is incorporated in the RGB signal is lost, although present. These findings demonstrate that OpenPose, while very useful, may not be powerful enough to accurately extract features for sign language, and that starting with raw RGB data could indeed provide additional value for the creation of accurate translation models.

The authors of the recent work are B. M. Chethana Kumara and H. S. Nagendraswamy from University of Mysore who performed numerous experiments to compare the performance of the proposed sign language recognition method [8]. They used a Canon 600D camera to collect a large number of ISL videos, containing different signs executed by Indian native signers. As practiced in the formation of the German SL dataset, additional videos were recorded from deaf students of various schools across the Mysore region, comprising of a total of 1040 videos of 26 signs of which are sign repetitions.

Training and testing percentages that were used in their experiments included repeating trials on 50 random sets of samples. They used the F-measure for the recognition performance analysis of their method because it consists of the precision and recall factors. They had an average of 84.81% recognition accuracy among their achieved goals using a 60:40 was used for the ratio of training to testing samples, the number of representatives involved were 434. They also found out that number of representatives and training samples impacts recognition rates of the signs high inter class varying signs being recognized better. This observation implies that there must be adequate representation to enable the identification of various forms of signs in their sample. Their outcomes show that a well-established framework is sufficient to obtain a high accuracy rate when identifying ISL signs in different situations.

More recently, Amit Moryossef along with Google, University of Surrey, and University of Zurich, has done a performance analysis on pose estimation models on sign language recognition (SLR) [9]. For the pose estimation, both teams used OpenPose and Holistic MediaPipe with highest validation accuracy ranging between 80 and 85 percent hence implying non-trivial errors and or implementation problems. Their pose estimation-based systems have better results than pretrained image feature extractors, although the latter had an accuracy of (38-68%), proving their better generalization towards completely novel signers and background contexts. Focusing on the official challenge evaluation, it is noticed that the test set accuracy of Team 2 combined system that employs OpenPose and

Holistic pose estimations elevates to 81.93%, compared to 78.35% for each of the pose estimation systems developed for each individual. This goes a long way in showing how incorporating the pose estimation features will enhance the general performance of an SLR camera.

Researchers Sang-Ki Ko and team from Korea Electronics Technology Institute (KETI) created a sign language translation model that they tested using the said KETI dataset for performance examination. Integrated in PyTorch, their model was trained using the Adam optimizer for 50 epochs with the learning rate being adapted at epochs 20, 40 [10]. They used dropout regularization, gradient clipping and many normalizations to improve the performance. The team compared different approaches by the given standard measures such as accuracy, BLEU, ROUGE-L, METEOR, and CIDEr, where the text was trained in the sentence level and the gloss level. They also considered different impact of the feature normalization, attention concepts, the data augmentation parameters and the number of frames being sampled. To this end, their findings stressed on the significance of sound training approaches and proper normalization procedures in enhancing translation precision and speed, which would be beneficial for optimizing live sign language translation systems.

Laia Tarres, Gerard I. Gallego, Amanda Duarte, Jordi Torres, Xavier Giro-i-Nieto, and other their colleagues from Universitat Politecnica de Catalunya, Barcelona Supercomputing Center, and Amazon performed accurate performance evaluation of the Sign Language Translation (SLT) models by employing the BLEU score which is widely used in machine translation [11]. They employed this via sacreBLEU and also introduced the reduced BLEU (rBLEU) to pay special attention to semantically relevant words, especially beneficial in cases of lower translations. They were able to show through their experiments that their optimized Transformer model run gave better quality in translations. The team then noticed the model rated shorter sequences better and struggled with longer sequences, which is consistent with the authors' observations of how the model fails to capture the signed videos' meaning fully. They employed a flexible grid search for hyperparameter tuning of the neural model, and experimented with the preprocessing of the text data which contributed to raising the rBLEU score while applying lower casing improved the rBLEU scores showing how important the preprocessing of text data in order to enhance the performance of the model is.

The recent work was done by Garrett Tanzer, Maximus Shengelia, Ken Harrenstien of Google and David Uthus of Rochester Institute of Technology on Sign language translation and the effect of context. It was essential for them to decide the translation quality, so they applied quantitative metrics like BLEU and BLEURT [12]. The study established that on a sentence-by-sentence level translations the BLEU score was 19.8 which afterward somewhat rose to 21.5 with additional context. However, annotators found to exist a modest numerical progression that was considered significant, 33.3% of the cases needed further information to be able to increase the context of the given sentence completely. Some of the challenges experienced were, indeterminacy of the referent, grammatical faults, speedy fingerspelling, and difference in dialect. The human baseline metrics were also different from each other and depended on the approaches of the interpreters used in translations that were assessed at varying levels, fluctuating between 5.2 BLEU to 39.5 BLEU. This paper stresses the significance of the context in translating the sign language and also the difficulties in such process for both human and machine.

How2Sign dataset was translated by Xavier Giro Nieto, Laia Tarres Benet and Patricia Cabot Alvarez and they used the Transformer based model for doing the translation [13]. They focused on the issue of translation performance and used the BLEU scores for the grading of the output. Even though for recognition their model has used the English sentence instead of the gloss annotations, their main goal was oriented on achieving the better result for this translation. They compared their work with the findings of previous studies using the dataset called PHOENIX2014T, however, they analysed recognition employing the Word Error Rate (WER) and translation using Bilingual Evaluation Understudy (BLEU) rates with Reciprocal Overlapping of Necessary Representatives (ROUGE) and Character n-Gram Frequency (CHRF). It is also beneficial as it provides them with a comprehensive formula on how to evaluate the performance of their model concerning the SLS translation tasks.

The authors Pooya Fayyazsanavi, Antonios Anastasopoulos, and Jana Košecká managed to achieve high performance in Gloss2Text translation through the usage of LLMs, data augmentation, and label-smoothing loss function [14]. From here, Their method indicated a 3.75% relative improvement in BLEU-1 and 6.69% in BLEU-4, 5.38% in ROUGE and 2.07% in CHRF++. For data augmentation, they employed paraphrasing and back translation to improve the model's relativity. They also checked for overfitting in large models and also employed LoRA techniques to counter the problem. Their proposed idea of smoothing labels really enhances generalization and narrow down the difference in performances between a development set and a test set. Altogether, their method did not only increase the accuracy of translating material but also boosted model effectiveness.

### 2.3.2 IOT based approaches

UNS students have created a sign language translator called "Glova", the glove which would help people who have communication and hearing issues [15]. These devices are worn on the fingers using flex sensors that transform the hand movements into textual and audible inputs for improvement of communication. To ensure that there is an accurate text correction to improve the results, the system uses the Levenshtein distance algorithm, which does not have a gesture similarity. Even in its early usage, It shown an improvement in accuracy of the real-time translation making it a worthy tool in promoting inclusion.

UCLA researchers Angeles created a glove that in real time translates ASL for the hearing-impaired into spoken English [16]. With stretchable sensors, the glove is designed to monitor hand's motion in a way that translates it into electrical signals which are sent to an app contained in a smartphone. This system also employs adhesive sensors particularly in recognizing the facial expressions which in turn helps to increase the level of translation quality. On the 660 hand gestures, they said it passed the experiment with a success rate of 98%. 63% recognition rate. The described approach makes it possible to achieve sign language and non-sign language users' interaction, which proves its applicability across multiple scenarios.

### 2.3.3 Hybrid approaches

Tuning an A team assignment based on the performance analysis derived from the B. Sc in Computer Science University of Colombo is to establish the applicability of different machine learning models in identifying Sri Lankan sign languages [17]. The researchers were used five classifiers which they named as Logistic Regression (LR), Linear Discriminant Analysis (LDA), Ridge Classifier (RC), Random Forest Classifier (RFC), and Gaussian Naïve Bayes (NB). In this case, the classifier that performs the best average 10-fold cross-validation accuracy of about 76.18% is the LDA. The result shows that this model is the most stable model out of all the models.

Precision, recall, and the F1-score were used to measure the correctness of the LDA classifier. The average precision was 0.81, recall was 0.79 and the F1-score was 0. 79 that still proved to be fair when measured. Further, the performance measure was interpreted using confusion matrices and received operating characteristic (ROC) curves, the majority of classes attained high accuracy and under the curve (AUC) scores.

Possible effects of some of these approaches, such as Principal Component Analysis (PCA) and Singular Value Decomposition (SVD), were considered. Further analysis of the results reveals that the LDA model was able to sustain cross validation accuracy even with a lesser selection of features, with accuracies reaching 75.78 % with 40 features, 78.07 % of 80 total amount of features of the face and 79.16% with 100 features respectively.

R. M. Rishan, S. Jayalal, and T. K. Wijayasiriwardhane from the University of Kelaniya evaluated their sign language recognition system by analyzing three types of errors, deletion(D), insertion (I), and substitution (S) [18]. They implemented a system that got an 80% clinical detection ratio and 0.1% Word Error Rate (WER) for static, multiple meaning signs which they claimed had high reliability because of their static nature. For dynamic combined signs, the detection ratio was measured to be 70 % with WER of 12 % and the reliability coefficient is 0.63 %. In general, according to the results, the described system has the average level of accuracy-the static signs for 80% and 77% for the dynamic signs for all 40 participants.

Two authors Eak-Wei Chong and Boon-Giin Lee of Keimyung University have used performance analysis on their ASL recognition model employing SVM as well as DNN classifiers consisting six different feature sets [19]. In their study, they found that the DNN reached better accuracy than the SVM by 90.58%  main accuracy for 26 ASL letters and 85.65% secondary accuracy at the maximum level for 36 classes (letters and digits) compared to the results of SVM equals to 75%. 5% and 67. 54%, respectively. It is therefore evident that the feature group C6 provided the highest mean accuracy rate of 83.78%. The C6 feature group was the most promising for correct classification as it provided the highest accuracy of the classifier which was 93. 81% for 26 classes and 88.79% for 36 classes. Again, for the SVM classifier, C4 was slightly better than C6 for the 26 classes but C6 was the best for the 36 classes. Overall, when using the mentioned feature groups, it is possible to state that using the DNN based on the six composite feature group gives the highest recognition rates for ASL.

To compare the results of the proposed sign language recognition strategy, Jordan J. Bird, Anikó Ekárt, and Diego R. Faria performed experiments [20]. Thus, based on the VGG16, they had increased the accuracy to the best value of 88.14% which was established using a layer of 128 neurons. When put together with the CNN and Leap Motion models into what they call the multimodality network, they were able to achieve a impressive 94.44% accuracy. This fusion proved to be more accurate than the individual models, the CNN on its own recorded an accuracy of 88014% while the Leap Motion model predicted, 72.73%. Also, the use of the leave-one-subject-out validation strategy demonstrated that the multimodality model’s average accuracy was 92.12% accuracy. In transfer learning, they also discovered that the preferred model was the multimodality model trained with BSL weights followed by ASL, which shows that data insufficient influences the performance of the multimodality model.

# Chapter 3: Methodology

## 3.1 Methodology

- Study of Sign Language Basics and environment
    - Comparative Analysis of Sign Languages.
    - Field Visits to Deaf Schools.
- Finalize the use case
- Selection of Approach
- Review of Existing Projects
    - Existing sign language recognition systems
    - Existing Tele Health Systems
    - Evaluate Data Availability and Suitability
- Technical Preparation
- Data Acquisition
- Data Preprocessing and Augmentation
- Model Development
- Model Evaluation and Improvement
- Mobile App Deployment

## 3.2 Detailed Methodology

We are planning to continue our research work as follows,

### 3.2.1 Study of Sign Language Basics and environment

Learn the fundamentals of sign languages, to clarify the following problems,

- how deaf people communicate
- which body parts are used to Communicate themselves
- how different sign languages are differ from each other, and how people learn them.

- identify the relationships among ASL, BSL, and SLSL.
- look into the difficulties deaf persons have interacting with the community.

Visit deaf schools and talk with sign language-speaking teachers to verify the availability of local sign language data and current real-world mobile applications.
By reading papers and publications, learn about the various approaches currently in use for sign language recognition and explore their benefits and limitations. Additionally identify the approaches to get around the current drawbacks.

### 3.2.2 Selection of Approach

By referring to papers, identify the main approaches and evaluate their advantages and limitations.

- Vision based Approach
- IOT Based Approach
- Hybrid Approach

According to the advantages and drawbacks identify the most suitable approach to our case.

### 3.2.3 Finalize the use case

Finalize The use case and specify our Scope that we can get started. Analyse various possible uses, including,

- Telemedicine
- Banking services
- Public transportation
- Hospital systems

And decide which is best for our use case. For selecting the best use case Need to consider about several key factors.

- Community Impact
- Practical Usage
- Data Availability
- Technical Feasibility

### 3.2.4 Review of Existing Projects

Study the existing sign language recognition systems, usability and limitations of them. And get know about the Technology behind the projects.

Study existing telemedicine systems and the treatment procedures in hospitals to gain a better understanding of the challenges faced by the deaf community when accessing these services. This can be achieved by meeting with medical professionals (Doctors, Nurses, Medical officers) and referring to relevant papers.

Check the data availability and suitability for our task, and identify the best dataset to initialize model building.

### 3.2.5 Technical Preparation

Based on the papers read and the projects reviewed, identify the key technologies and gain practical experience to start the project. Mainly Need to focus on Followings.

- CNN
- Image Processing
    - Open CV Library
    - Mediapipe Library
    - PIL Library

    - Open Pose
- GRU
- LSTM
- TRANSFORMERS
- Mobile Application Development Related Technical Knowledge

### 3.2.6 Data Acquisition

Examine the different types of data and their usage. Evaluate which kind of data is usable for our task.

- **Image Data**

  Image data consists of static images capturing individual signs or gestures. Compared to video data, this kind of data requires less storage space and processing power, making it simpler to prepare and maintain. But because it doesn't have temporal context, it isn't as good at identifying continuous sign language or sentences. Best for isolated sign recognition or fingerspelling where temporal context is not critical.

- **Word Level Video Data**

  Word-level video data consists of brief video clips that show specific words or moving signs. The temporal context that this data type offers is crucial for capturing the dynamics of each sign. Compared to static images, word-level video data can be used to recognize and understand the transitions between hand forms and movements for particular words. word-level video data might not fully convey the content of a sentence, which could result in misunderstandings when signs have varied meanings depending on the situation.

- **Sentence Level Video Data**

  Longer video clips that capture complete sentences or sign sequences make up sentence-level video data. With the ability to capture the complete dynamics and transitions between signs in a sentence, this data type offers extensive temporal context. Sentence-level video data is more useful for real-world applications where continuous sign language recognition is needed since it is more suited for comprehending the grammatical structure and context of sign language.

- **Depth Data**

  Depth data is captured from depth-sensing cameras, such as Kinect, providing three-dimensional information about hand and body movements. This data type captures 3D spatial information.

  Depth data is adaptable to changing lighting conditions and backgrounds and can distinguish between comparable signals based on depth and spatial location. Depth data, however, needs specific hardware, which might not be readily accessible. Because 3D information must be processed, it requires more computer power and is more difficult to preprocess and analyze than 2D picture or video data.

### 3.2.7 Data Preprocessing and Augmentation

Data preprocessing and augmentation is a important step to improve the model performance. For Image and video data the following preprocessing steps are
Crucial,

- **Data Cleaning:** Removing noise, handling missing values, and correcting errors in the data to improve data quality.
- **Data Normalization:** Scaling the data to ensure all values fall within a specific range, which aids in improving model convergence and performance.

- **Feature Selection:** Identifying and selecting the most relevant features from the data to reduce dimensionality and enhance computational efficiency.
- **Image Augmentation:** Generating additional images or frames by applying transformations to existing data, thereby improving model generalization and robustness.

Video is a sequence of image frames; therefore, these preprocessing steps are equally applicable to video data to enhance the model performances.

The lighting and background of picture frames affect the performance of the model and pose problems for current methods; to mitigate these issues, we propose preprocessing steps that include drawing the skeleton plot and removing the background before feeding the data to the model. Study the MediaPipe and OpenPose libraries and Optical Flow Analysis. Prepare a Numerical data set by taking the optical flow vectors of each video files to efficient usage of the computation resources.

Optical Flow Vector Analysis, illustrated by Figure 1, tracks movement of objects from frame to consecutive frame in a video. The red blurred circles, numbered from 1 to 5, indicate the progressive motion of a tracked point over time, and the white arrow indicates the estimated motion trajectory.

Optical flow vectors measure the displacement of pixels in a frame, and this measures also indicate their respective direction and velocity. Particularly in applications such as sign language recognition, tracking hand and body movements is very important. An extraction of optical flow vectors would enable computational models to interpret motion patterns better, leading to improved accuracy in gesture-based systems.

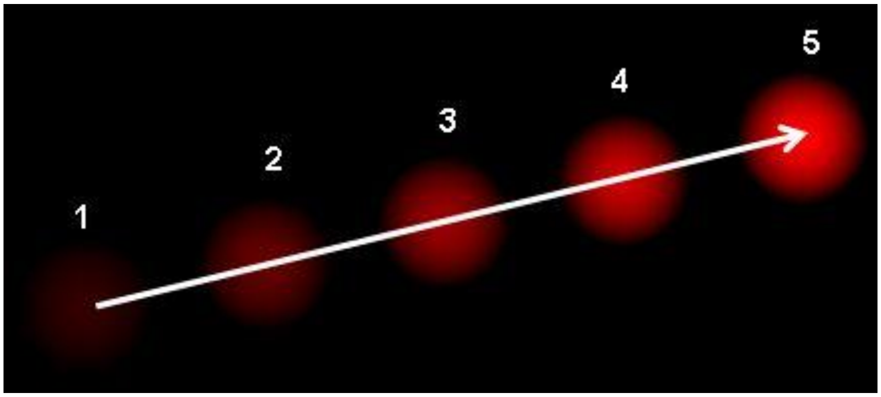


**Figure 1 : Optical Flow vector Analysis**

Optical Flow Vector calculation, illustrated with the diagram in Figure 2, refers to the method of determining the motion of objects or pixels from one frame of a video to another. The figure shows movement from the point of origin P(x,y,t) to a new position P(x+Δx,y+Δy,t+Δt), therefore establishing pixel movement in time down both the x as well as y directions.

The equation mentioned is referred to as the Optical Flow Constraint Equation to be used in estimating motion vectors. The estimation of motion vectors from the constraint equation ultimately allows motion tracking, gesture recognition as well as video sign language translation systems, among many others, to have real-time applications.

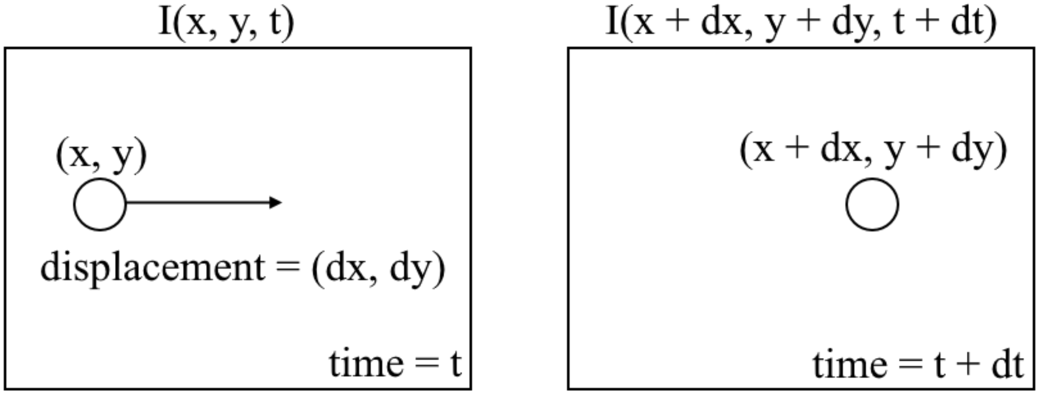


**Figure 2 : Optical Flow vector Calculation**

Optical Flow vectors can represent the velocity and the movements of specific pixel sequence of the image frames.

The figure 3, provides an example of image augmentations in use in data pertaining to sign languages. The augmentation technique takes a given skeletal representation and causes it to yield an abundance of variances in order for the models to be robust and generalize better. This, in turn, helps counter the effects of overfitting through diversity in training samples.

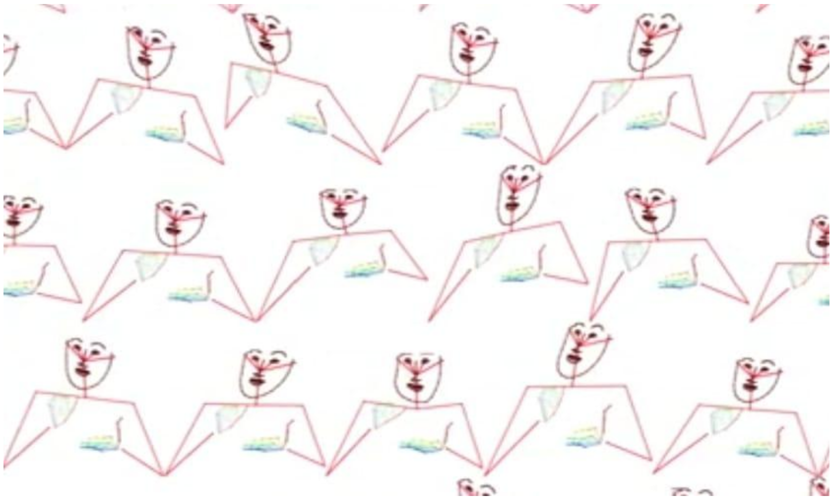


**Figure 3 : Image Augmentation**

This figure 4, highlights a skeletal representation of human gestures with emphasis on key points of the face, hands, and upper body. The key points extracted and shown in colour are applied in motion tracking and sign language recognition. This visualization, therefore, helps in capturing hand and facial dynamics, both key for accurate sign language interpretation and gesture-based interaction.

This figure 5, illustration shows a stylized skeletal representation of a human performing a gesture. The model highlights important facial, hand, and body structures, with distinct colour gradients on the hands to emphasize finger position and movement. This visualization is applied in gesture recognition and sign language translation in order to help in analysing the motion patterns and hand articulation.

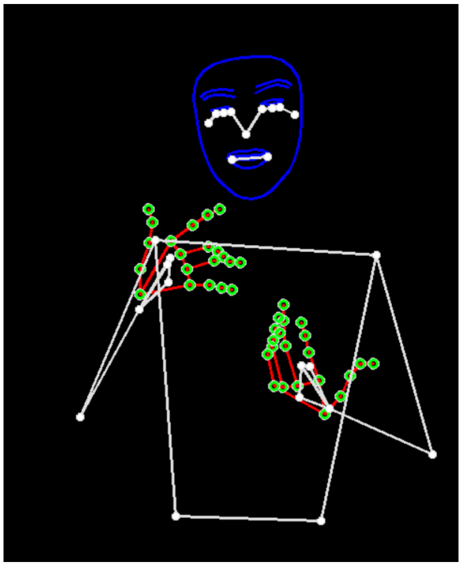


**Figure 4 : Skeletal Representation of Human Gesture Key Points**

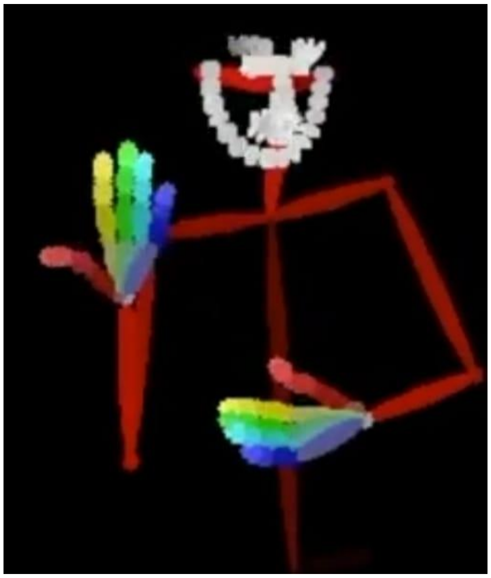


**Figure 5 : Stylized Skeleton Figure**

### 3.2.8 Feature Extraction

Feature extraction forms a critical part of sign language recognition systems wherein raw video input is converted to syntactic representations interpretable by a machine learning model. For this task, the I3D (Inflated 3D ConvNet) model is employed for feature extraction which has been established to perform well in video comprehension activities.

The I3D model leverages 2D convolution kernels that are expanded into 3D, making it possible for I3D networks to extract spatial and temporal elements within a video sequence. Consequently, this technique enables the model to analyse sign language videos in three dimensions allowing for spatial and temporal sign language patterns to be analysed.
Here are Some of the Key Steps:

- **Video Data Preprocessing**: This process involves scaling and normalizing of the video frames so that they conform to the input for the I3D model.
- **Feature Extraction**: Through this process, the I3D model treats the video frames as an input stream consisting of motion and appearance visuals and spectrally analyses each segment into a concise high-dimensional feature vector.
- **Dimensionality Reduction**: Appropriate algorithms such as PCA can also be used to cut down the number of features utilized for a downstream process as long as the critical information is not compromised.

Physical space and feature reward structure are crucial aspects of machine learning models. Therefore, I3D features that have been extracted allow us to connect the visual fractals with the respective reward.

This figure 6, shows the general architecture of the I3D (Inflated 3D) network for action recognition in videos. The model utilizes two streams, the RGB stream and the optical flow stream, in parallel. The former processes the raw color frames, while the flow stream processes data on motions derived using the TV-L1 optical flow algorithm. The two streams are combined on their features and forwarded through a softmax layer for the final prediction of the action. This dual stream architecture captures the spatial and temporal dynamics of video data, thereby improving the recognition accuracy.

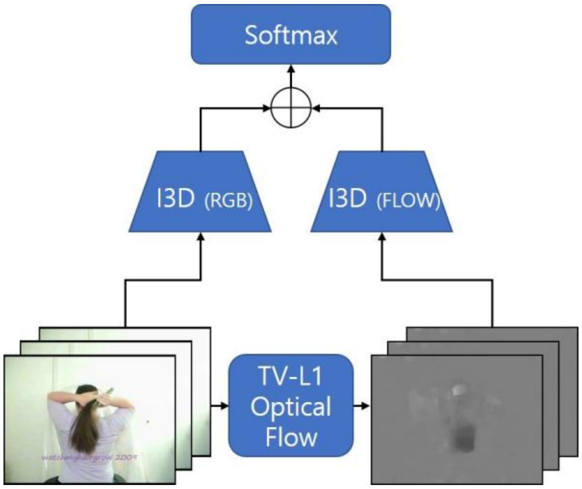


**Figure 6 : Overall Architecture of I3D**

### 3.2.9 Model Development

With the pre-gathered knowledge, build different architectures and train models using the selected dataset.

- GRU Architecture
- LSTM Architecture
- Transformers Architecture
- LSTM + Transformers Architecture

Continuous sign language recognition is a time-series sequential data analysis task. Therefore, GRU, LSTM, and Transformer-based models are primarily aligned with this approach.

**Table 1 : Comparison of GRU, LSTM, and Transformer-Based Models for Sign Language Translation**

<table>
<tr><th>GRU</th></tr>
<tr><td>

computationally efficient and effective for sequential data modelling.

Might struggle to retain long-term dependencies in very long sequences.

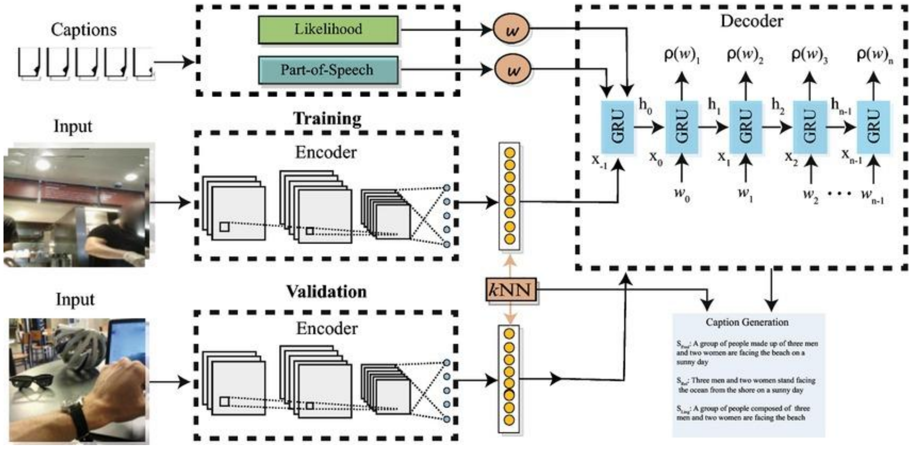

Captions
Likelihood
Part-of-Speech
w
w
Decoder
ρ(w)₁
ρ(w)₂
ρ(w)₃
ρ(w)ₙ
Training
Input
Encoder
GRU
h0
h1
h2
hn-1
x-1
x0
x1
x2
xn-1
w0
w1
w2 ··· wn-1
kNN
Input
Validation
Encoder
Caption Generation
S: A group of people made up of three men and two women are facing the beach on a sunny day
S: Three men and two women stand facing the ocean from the shore on a sunny day
S: A group of people composed of three men and two women are facing the beach

**Figure 7 : Overall Architecture of GRU model**

During training, GRUs update their hidden states at each time step to capture temporal patterns in video frames, learning the mapping between input sequences and corresponding text translations,

the model processes the feature sequence frame by frame, predicting output tokens based on the learned context.

</td></tr>
<tr><td>LSTM</td></tr>
<tr><td>

LSTM introduces a memory cell and gates (input, forget, output) to handle long-term dependencies and prevent vanishing gradient problems.

More computationally intensive than GRU

</td></tr>
</table>

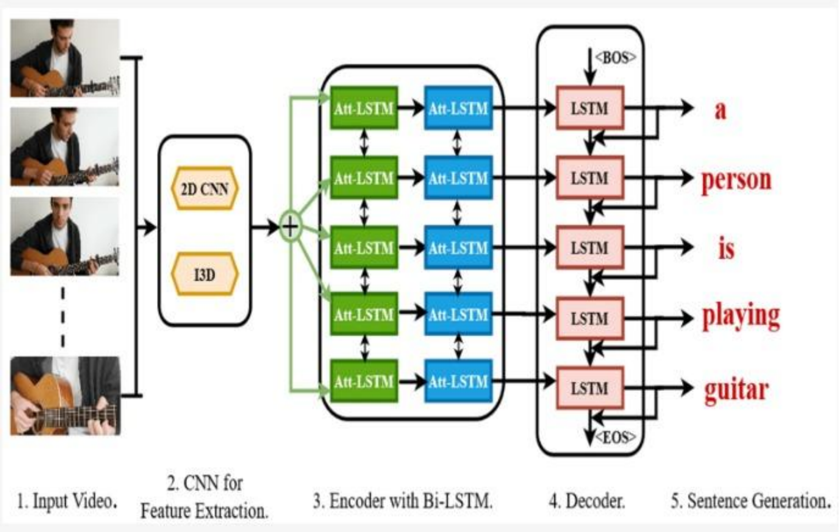


**Figure 8 : Overall Architecture of LSTM model**

The LSTM processes video features frame by frame, updating its hidden and cell states to retain important temporal patterns. It learns to map input sequences to text translations.

During inference, the LSTM processes the sequence of features, predicting output tokens at each step. These predictions are then decoded into a full text sequence.

TRANSFORMERS

Transformers rely on self-attention mechanisms to model relationships across all elements in a sequence simultaneously

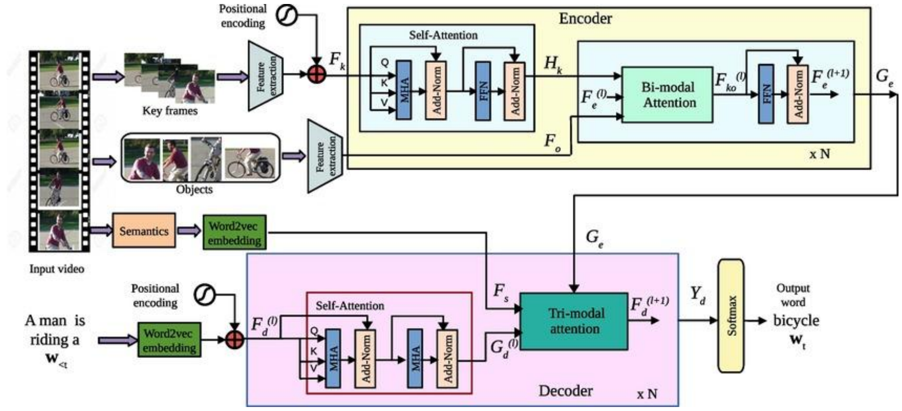


**Figure 9 : Overall Architecture of Transformer model**

Transformers process all frames of I3D optical flow features simultaneously, learning relationships between frames regardless of distance. The model uses positional encodings to understand the sequence order and learns to map features to text translations.

The Transformer processes the entire sequence at once, generating output tokens by attending to all frames and decoding them into a complete text sequence.

The model that best aligns with our case depends on several factors, such as dataset size, quality, variations, response time, and complexity. Model performance can be evaluated using metrics like BLEU score, WER (Word Error Rate), response time, etc.

### 3.2.10 Model Evaluation and Improvement

Evaluate the model performance and Explore ways to further improve accuracy of the model.

We suggest using the following metrics and factors to assess the model's performance.

- Word Error Rate (WER) - Measures the percentage of words in the predicted sentence that differ from the reference sentence.
- Sentence Error Rate (SER) - Measures the percentage of sentences that are incorrectly predicted compared to the reference sentences.
- Accuracy - Measures the overall correctness of predicted sentence
- Latency - Measures the time taken by the model to process and recognize a sentence of sign language in real-time.
- BLEU Score (Bilingual Evaluation Understudy) - Assess quality of predicted sentences by measuring against other sentences and focuses on the precision of n-grams, the blue score is a metric that is popular due to it assessing the semantic and syntactic similarity of translated languages within natural language processing tasks. This metric facilitates the calculation in a systematic way of how comparable the predicted sentences are to the reference sentences based on their lexical and contextual meaning. The BLEU score has its great use when evaluating the translation of sign language at the sentence level, because an account of the language's grammar or syntax would be also taken.

These metrics above together ensure that the effectiveness, the efficiency, and the real-world value of the model are evaluated, and areas of improvement are pinpointed.

### 3.2.11 Mobile App Deployment

With the best model, deploy the pipeline and the mobile app to use the model on an edge device. (With considering the better cloud platform)

During the model deployment stage, Need to Consider real-time communication protocols to minimize latency in communication. And Use techniques to build the application in light weighted for end users.

This figure 10, explains the patient-doctor communication system, comprising sign language input and voice translation. The system allows a patient to express his or her symptoms through signs before the front camera of a mobile device, which captures and encodes video. This is then sent for processing where the input is decoded and converted to text, which can be displayed on both patient and doctor application interfaces, with an optional text-to-voice output. Doctors can respond by activating the microphone, and their voice input is then encoded, converted to text, and displayed to the patient for unimpeded communication.

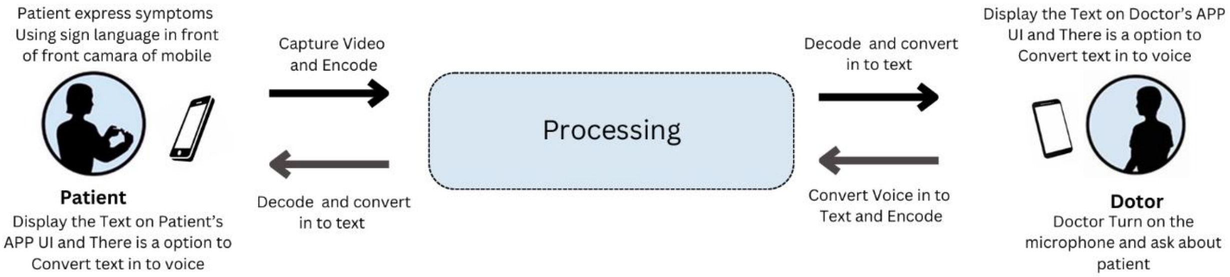

Patient express symptoms Using sign language in front of front camara of mobile
Patient
Display the Text on Patient's APP UI and There is a option to Convert text in to voice
Capture Video and Encode
Decode and convert in to text
Processing
Decode and convert in to text
Convert Voice in to Text and Encode
Display the Text on Doctor's APP UI and There is a option to Convert text in to voice
Dotor
Doctor Turn on the microphone and ask about patient

**Figure 10 : Proposed System**

The below figure 11, shows the suggested architecture for the video communication system that employs ML techniques. At the beginning of the pipeline, some video data at the sentence level is subject to data cleaning and preprocessing, which involves scaling, resizing, augmenting, and modifying backgrounds. Features are extracted from continuous videos to build and evaluate models with LSTM, and Transformer architectures. The trained model is stored and integrated into the cloud platform so real-time communication can be maintained between the peer and UDP protocols. Real-time prediction is enabled through the cloud function so that interaction through video communication can also be supported with text translation on the mobile UI.

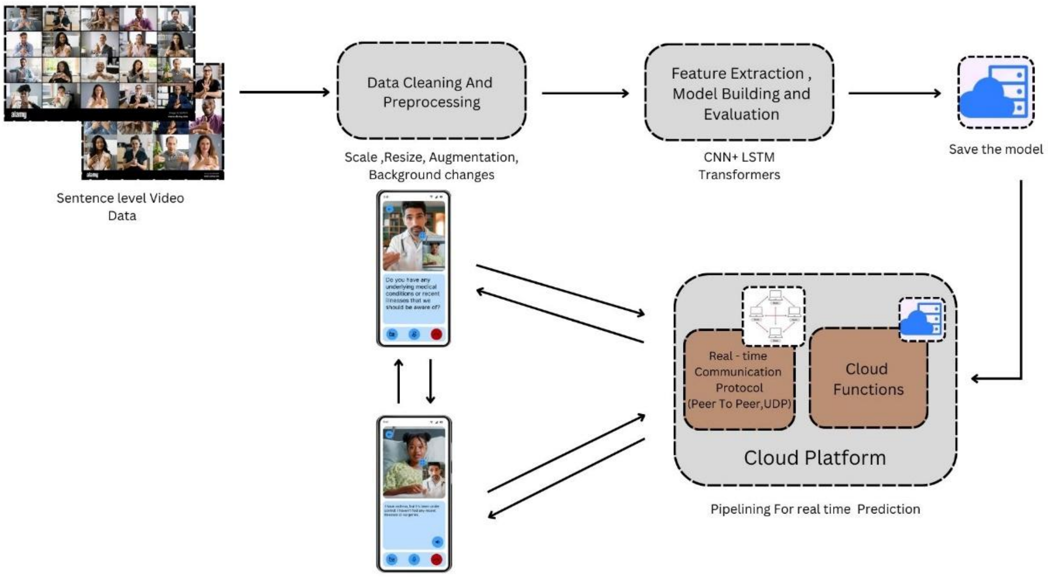

Sentence level Video Data
Data Cleaning And Preprocessing
Scale ,Resize, Augmentation, Background changes
Feature Extraction , Model Building and Evaluation
CNN+ LSTM Transformers
Save the model
Do you have any underlying medical conditions or recent illnesses that we should be aware of?
I have asthma, but it's been under control. I haven't had any recent illnesses or surgeries.
Real - time Communication Protocol (Peer To Peer,UDP)
Cloud Functions
Cloud Platform
Pipelining For real time Prediction

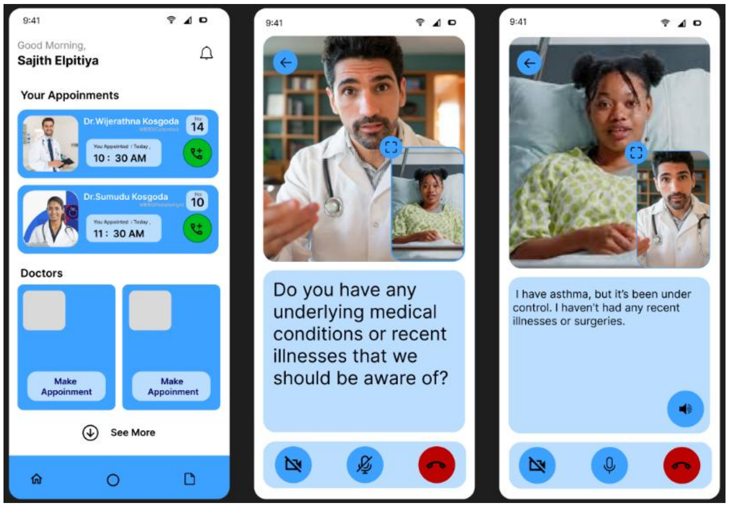
**Figure 11 : Proposed Architecture**

We propose a mobile application to facilitate communication between doctors and deaf patients. Using Peer to Peer Network protocol And the application initiates a video call between doctor and deaf person, and displaying the video feed on the top half of the screen and a text display widget on the bottom half Text field display the translation of sign language.

1. Doctor's UI: The doctor asks questions orally, and the app converts speech to text real time, displaying it on the patient's text widget.

2. Patient's UI: The patient communicates using sign language, The app converts to English text and display it on doctor's text widget.

The telehealth application for mobile consultations is presented in Figure 12. The first screen displays the user's scheduled appointments with all relevant details of their doctor and appointment times. The screen also allows the user to initiate a video call or access further information. The second screen shows a video consulting session in progress, with doctors interacting with their patients via video and text questions. The third screen shows a text reply from the patient, further enhancing communication for the hearing- or speech-impaired. Accessibility, usability, and seamless doctor-patient interaction were the major design drivers for the UI.



Do you have any underlying medical conditions or recent illnesses that we should be aware of?

I have asthma, but it's been under control. I haven't had any recent illnesses or surgeries.

**Figure 12 : Proposed Mobile Application UI**

# Chapter 4: Experimental Framework

## 4.1 Introduction

This chapter presents the experimental framework established to design, develop, and evaluate our sign language translation system for telehealth applications. The framework outlines the systematic process followed, from reviewing existing research to implementing and testing our proposed models in real-time communication scenarios.

We began by conducting an extensive literature review to understand the current technological landscape in sign language recognition. This review helped in identifying key methodologies, datasets, and limitations of existing systems, guiding our decision to adopt a computer vision-based approach. Building on this foundation, we explored several deep learning architectures - such as CNN, LSTM, GRU, and Transformer networks to determine the most effective model for our application.

The experimental framework further details the process of dataset selection, data preprocessing, model development, and performance evaluation. Both word-level and sentence-level datasets were examined to identify the most suitable data representation for real-time translation. Additionally, the integration of a fine-tuned Large Language Model (LLM) was explored to enhance grammatical correctness and contextual understanding of translated sentences.

To ensure the practical relevance of our system, we also engaged directly with the deaf community to gain real-world insights into their communication challenges and expectations. Finally, the chapter concludes with the description of our two-way real-time implementation that demonstrates effective communication between a patient and a doctor using socket-based data transmission.

## 4.2 Experimental Framework

We explored research papers to understand the current technology for recognizing sign language. This review helped us identify the benefits and limitations of existing systems and select a computer vision-based approach for our project. Our focus then shifted to vision-based methodologies, specifically analyzing research on CNN, LSTM, GRU, and Transformer architectures. These investigations highlighted the advantages and drawbacks of each technology.

We refined our use case by referring to research articles about vision-based sign language recognition systems for telehealth. This process revealed several challenges and limitations:

- Variability in sign language gestures with different countries and regions.
- No enough data resources.
- Challenges in real-time processing and response.
- Output effected by the different lighting conditions.

Then we needed to select our dataset, considering two main options:

- Word-level video data
- Sentence-level video data

To select the most effective dataset type, we have referred additional papers comparing the performance of systems using sentence-level and word-level datasets. We also Read the use of LLMs to enhance output.

According to these papers, we recognized the need to work with different deep learning architectures and algorithms to fulfil our task. Therefore, we both navigated gain practical knowledge about this technical background.

There are lot of advantages and drawbacks of different technologies. By addressing these drawbacks, our research theme is to develop a service for individuals with hearing and speech impairments, enabling them to communicate comfortably and effectively to meet their medical requirements.

To gain real-world experience in communicating with deaf individuals and to understand sign languages and the difficulties they face when interacting with the general public, we visited two different deaf schools.

1. Nuffield Special School for Deaf and Blind Students (Kaithady)
2. Sri Sudarshi Special School for Deaf and Blind Students (Bandarawela)

During these visits, we have interacted with teachers proficient in sign language and students who are partially blind and know sign language. We asked questions to clarify our understanding and gather information we needed to know directly from those with lived experiences. And asked about video data availability for learn sign language.

**Figure 13 : Engaging with Hearing and Speech Impaired Students for Data Collection**

We have get known about the existing mobiles applications developed for deaf communities and functions of the applications.

- “Sanwaada” – Dictionary containing Sign language videos and their meaning in Sinhala
- “SLSL dictionary” - Dictionary containing Sign language videos and their meaning in Sinhala, Tamil and English

- “Learning Material Issued by the NIE” - Dictionary containing Sign language videos and their meaning in Sinhala, Tamil and English
- “Tamil Sign Language Learning App” – Yarl IT hub development A Quiz type sign language learning app In Tamil medium

we had to determine the best type of data for our real-time language translation task , we compared the two main types of available data types:

1. Word-level video data
2. Sentence-level video data

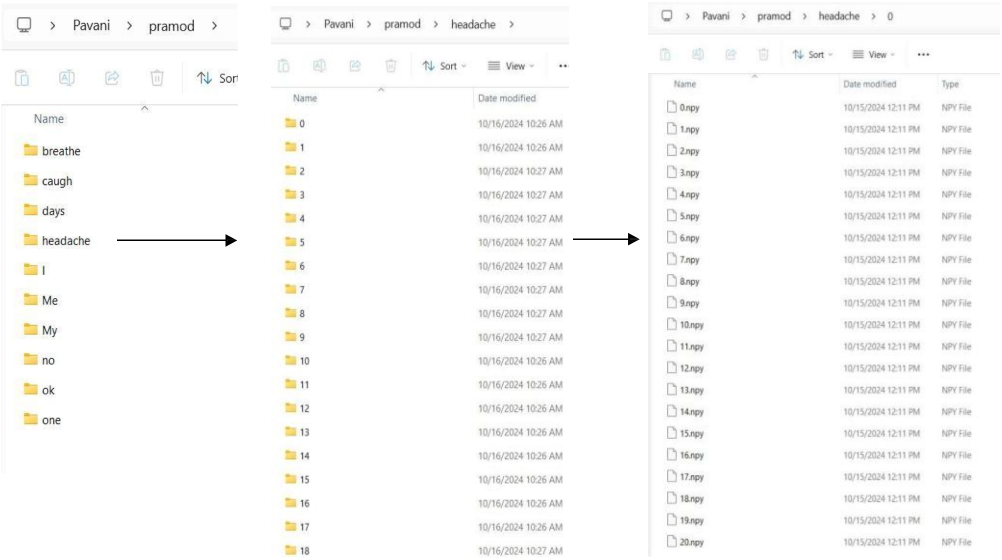
### 4.2.1 Initial Experiments

We began with our own word-level datasets and implemented a model using 10 telehealth-related British Sign Language (BSL) words. We referred to the BSL dictionary to find medical-related sign language words, practiced them, and recorded 200 videos for each word ourselves.

- Words: ‘I’, ‘My’, ‘Me’, ‘One’, ‘Breathe’, ‘Cough’, ‘Days’, ‘Headache’, ‘OK’, ‘No’.

Using a webcam, we recorded 200 videos per word. MediaPipe was employed to extract landmarks, generating NumPy arrays for 30 consecutive frames per video.

Therefore for 10 words, 2000 videos are used and prepared the data set as following structure.



| Name |
|---|
| breathe |
| caugh |
| days |
| headache |
| I |
| Me |
| My |
| no |
| ok |
| one |

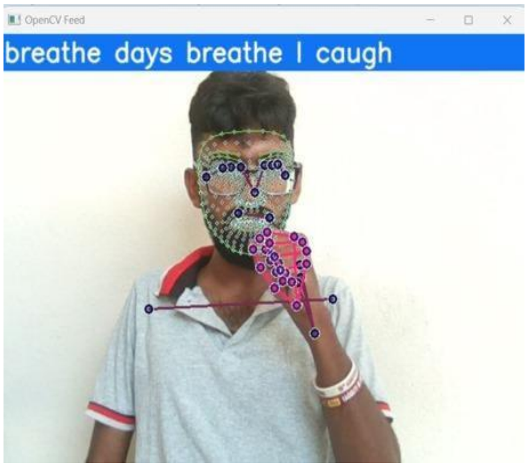
| Name | Date modified |
|---|---|
| 0 | 10/16/2024 10:26 AM |
| 1 | 10/16/2024 10:26 AM |
| 2 | 10/16/2024 10:27 AM |
| 3 | 10/16/2024 10:27 AM |
| 4 | 10/16/2024 10:27 AM |
| 5 | 10/16/2024 10:27 AM |
| 6 | 10/16/2024 10:27 AM |
| 7 | 10/16/2024 10:27 AM |
| 8 | 10/16/2024 10:27 AM |
| 9 | 10/16/2024 10:27 AM |
| 10 | 10/16/2024 10:26 AM |
| 11 | 10/16/2024 10:26 AM |
| 12 | 10/16/2024 10:26 AM |
| 13 | 10/16/2024 10:26 AM |
| 14 | 10/16/2024 10:26 AM |
| 15 | 10/16/2024 10:26 AM |
| 16 | 10/16/2024 10:26 AM |
| 17 | 10/16/2024 10:26 AM |
| 18 | 10/16/2024 10:27 AM |

| Name | Date modified | Type |
|---|---|---|
| 0.npy | 10/15/2024 12:11 PM | NPY File |
| 1.npy | 10/15/2024 12:11 PM | NPY File |
| 2.npy | 10/15/2024 12:11 PM | NPY File |
| 3.npy | 10/15/2024 12:11 PM | NPY File |
| 4.npy | 10/15/2024 12:11 PM | NPY File |
| 5.npy | 10/15/2024 12:11 PM | NPY File |
| 6.npy | 10/15/2024 12:11 PM | NPY File |
| 7.npy | 10/15/2024 12:11 PM | NPY File |
| 8.npy | 10/15/2024 12:11 PM | NPY File |
| 9.npy | 10/15/2024 12:11 PM | NPY File |
| 10.npy | 10/15/2024 12:11 PM | NPY File |
| 11.npy | 10/15/2024 12:11 PM | NPY File |
| 12.npy | 10/15/2024 12:11 PM | NPY File |
| 13.npy | 10/15/2024 12:11 PM | NPY File |
| 14.npy | 10/15/2024 12:11 PM | NPY File |
| 15.npy | 10/15/2024 12:11 PM | NPY File |
| 16.npy | 10/15/2024 12:11 PM | NPY File |
| 17.npy | 10/15/2024 12:11 PM | NPY File |
| 18.npy | 10/15/2024 12:11 PM | NPY File |
| 19.npy | 10/15/2024 12:11 PM | NPY File |
| 20.npy | 10/15/2024 12:11 PM | NPY File |

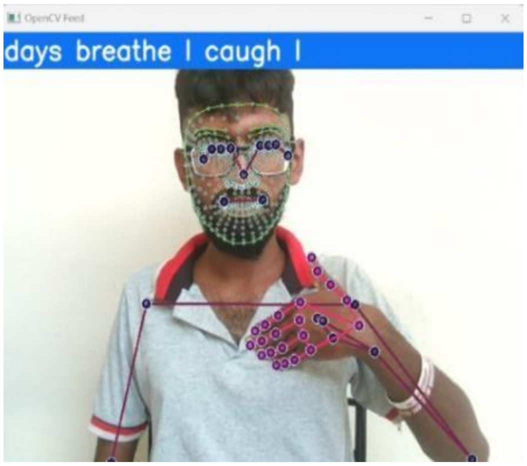
In this case, to make predictions with the sequential data, we planned to use an LSTM (Long Short-Term Memory) architecture.

Then , Compiled the model with 'Adam optimizer' and 'categorical cross entropy loss'. And trained the model with 50 epochs.

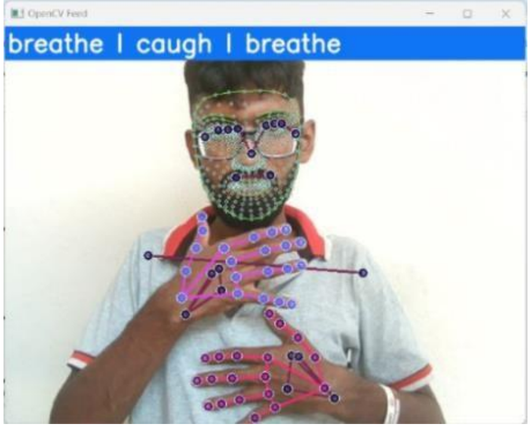
From that model we got some better predictions also Using Web Cam.

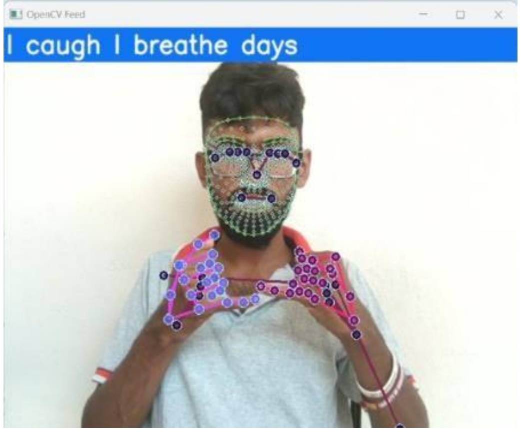


breathe days breathe I caugh



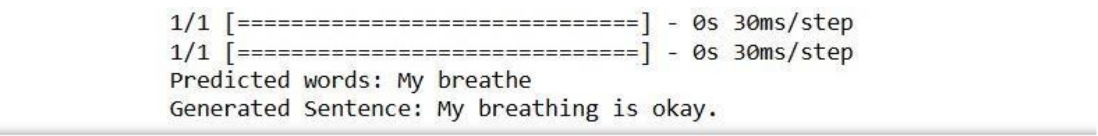
days breathe I caugh I





By only predicting this word sequence, we cannot generate grammatically correct sentences. Therefore, we have used a pre-trained LLM model, BART, with fine-tuning sentences that can be formed using the predicted words.

After Predicting the Word sequence then pass the word sequence through this finetuned LLM. Following is the sample predictions from the LLM.



```
1/1 [==============================] - 0s 30ms/step
1/1 [==============================] - 0s 30ms/step
Predicted words: My breathe
Generated Sentence: My breathing is okay.
```

Limitations of Word-Level Data,

- Limited contextual understanding.
- Poor grammatical accuracy.
- Requires advanced technologies like LLMs for continuous prediction.

### 4.2.2 Transition to Sentence-Level Data

Given the limitations of word-level data, we opted for sentence-level datasets, selecting HOW2SIGN due to its direct accessibility. This decision allowed us to start model building and experimentation promptly.

The selection process for the HOW2SIGN dataset was informed by several key issues which include the following,

- Direct Accessibility - Our workflow was made easier because incoming data acquisition was avoided as we were able to use the HOW2SIGN dataset which is publicly accessible, this means that methods of acquisition are not complex. This enhancement enabled us to begin model training and experimentation promptly and efficiently.
- Rich Multimodal Data - The addition of the strategy of combining learnings from multiple modalities helps bolster the ongoing development of robust models. The dataset features sign videos, audio recordings of people speaking, and English transcriptions with over 80 hours of parallel corpora that cater to supplementary materials of different mediums.
- Sentence-Level Annotations - Our projects' objective was to be able to convert universal sign language movements into coherent text, so honing in on the grammatical and contextual features of sign language phrases is integral to the sentence level annotation dataset we collected.
- 3D Pose Estimation Data - The lack of gloss annotations is overcome by adding 3D limbs and figures enabling models to convert signs in the new model into actions, and because HOW2SIGN is complimentary over these 3 hours of 3D pose estimation data, it helps in understanding spatial and temporal characteristics of signs.
- Variety in Vocabulary - A total of around 33,116 and 24,703 words comprises of HOW2SIGN's data vocabulary demonstrating the broad cross regional dialects it enfolds across various tutorial domains.

### 4.2.3 Model Development

We trained three models using the HOW2SIGN dataset, which contains 30,384 video samples with I3D feature extraction:

1. Encoder + T5 Decoder
2. Encoder + LSTM Layer + Decoder
3. Encoder + Transformer Layer + Decoder

After training all three models, they were evaluated on the train, validation, and test datasets. The evaluation metric used was the BLEU score. Below are the results for each model:

- Model 1: Encoder + T5 Decoder
    - Train BLEU Score: 0.0137
    - Validation BLEU Score: 0.0133
    - Test BLEU Score: 0.0129

- Model 2: Encoder + LSTM Layer + Decoder
    - Train BLEU Score: 0.0388
    - Validation BLEU Score: 0.0299
    - Test BLEU Score: 0.0312

- Model 3: Encoder + Transformer Layer + Decoder
    - Train BLEU Score: 0.0689
    - Validation BLEU Score: 0.0459
    - Test BLEU Score: 0.0478

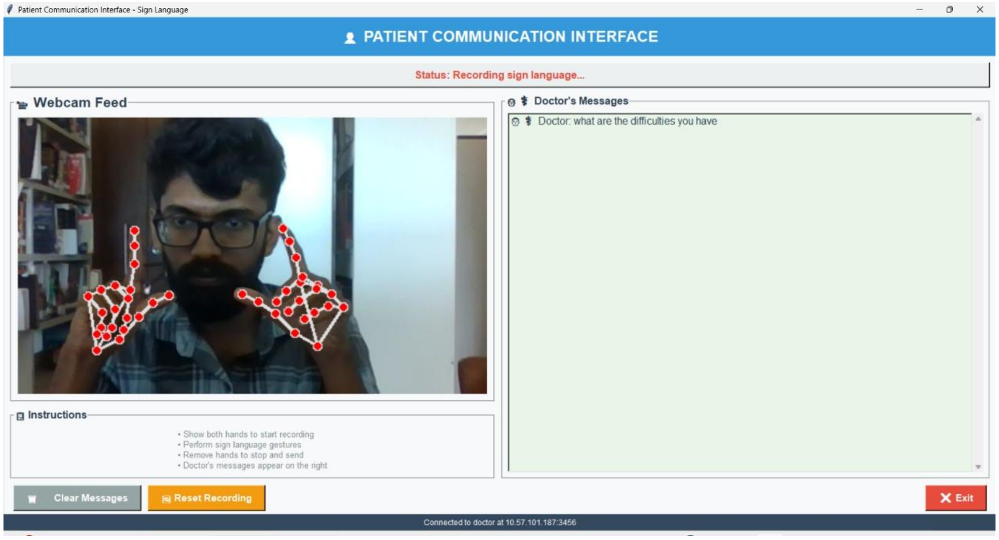
Evaluation using BLEU scores revealed that the **Encoder + Transformer Layer + Decoder** model performed best.

### 4.2.4 Real-Time Implementation

To demonstrate the practical application of the system, we implemented a two-way communication setup using socket programming. The setup includes the following components,

Patient-Side:

- Captures sign language videos via a webcam.
- Sends the captured videos to the doctor-side laptop using socket communication.


Patient Communication Interface - Sign Language
PATIENT COMMUNICATION INTERFACE
Webcam Feed
Doctor's Messages
Instructions
Clear Messages
Reset Recording
Exit

Status: Recording sign language...

Doctor: what are the difficulties you have

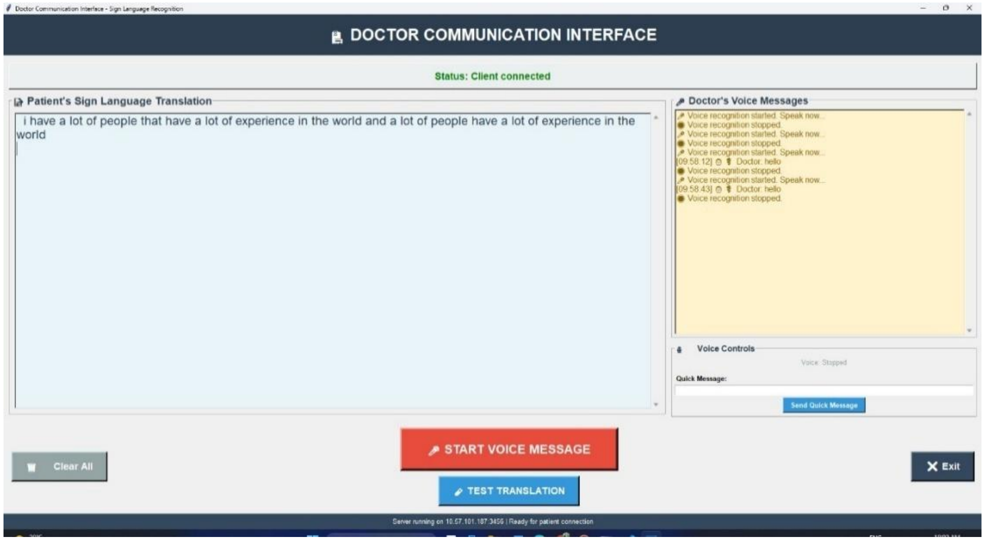
- Show both hands to start recording
- Perform sign language gestures
- Remove hands to stop and send
- Doctor's messages appear on the right

Connected to doctor at 10.57.101.187:3456

**Figure 14 : Patient-Side Sign Language Video Capture and**

Doctor-Side :

- Receives the sign language videos from the patient-side laptop.
- Extracts I3D features from the videos.
- Processes the extracted features using our trained model to recognize signs.
- Converts the recognized signs into text and displays the translations.


Doctor Communication Interface - Sign Language Recognition
DOCTOR COMMUNICATION INTERFACE
Patient's Sign Language Translation
Doctor's Voice Messages
Voice Controls
Send Quick Message
START VOICE MESSAGE
TEST TRANSLATION
Clear All
Exit

Status: Client connected

i have a lot of people that have a lot of experience in the world and a lot of people have a lot of experience in the world

```
Voice recognition started. Speak now...
● Voice recognition stopped.
Voice recognition started. Speak now...
● Voice recognition stopped.
Voice recognition started. Speak now...
[09:58:12] Doctor: hello
● Voice recognition stopped.
Voice recognition started. Speak now...
[09:58:43] Doctor: hello
● Voice recognition stopped.
```

Voice: Stopped

Quick Message:

Server running on 10.57.101.187:3456 | Ready for patient connection

**Figure 15 : Doctor-Side Sign Language Video Processing and**

The system supports voice-to-text communication from the doctor's side, enabling the doctor to send text-based messages back to the patient-side laptop.

This socket-based communication setup ensures smooth and efficient data transfer, enabling real-time interaction and bridging the communication gap between the patient and the doctor.

# Chapter 5: Experimental Results and Analysis

## 5.1 Introduction

This chapter presents the experimental results and performance analysis of the developed sign language translation system. The experiments were conducted to evaluate the effectiveness of different model architectures and to assess the system's performance in both word-level and sentence-level translation tasks. The objective of these experiments was to validate the feasibility of the proposed approach in achieving accurate and real-time sign language recognition and translation for telehealth communication.

The experimental phase began with the development of a Long Short-Term Memory (LSTM) model trained on a small, custom-created British Sign Language (BSL) subset consisting of ten medical-related words. This initial experiment was conducted to understand the model's capability to classify isolated signs and to evaluate real-time performance using a webcam-based setup. The model's predictions were analyzed using a multilabel confusion matrix to measure classification accuracy and identify commonly misclassified signs.

Building on the insights gained from the BSL subset, the research progressed to the HOW2SIGN dataset, a large-scale sentence-level dataset. The dataset's I3D (Inflated 3D ConvNet) feature extractions were used to train and compare multiple deep learning architectures.

Each model was evaluated using BLEU scores on training, validation, and test sets to measure translation quality. Comparative analyses between these architectures revealed that the Encoder + Transformer Layer + Decoder model achieved the best performance, demonstrating strong spatial-temporal understanding and sentence formation capability.

The following sections present detailed experimental results, accuracy comparisons, BLEU score evaluations, visual analyses, and model predictions that demonstrate the performance and practical applicability of the developed sign language translation framework.

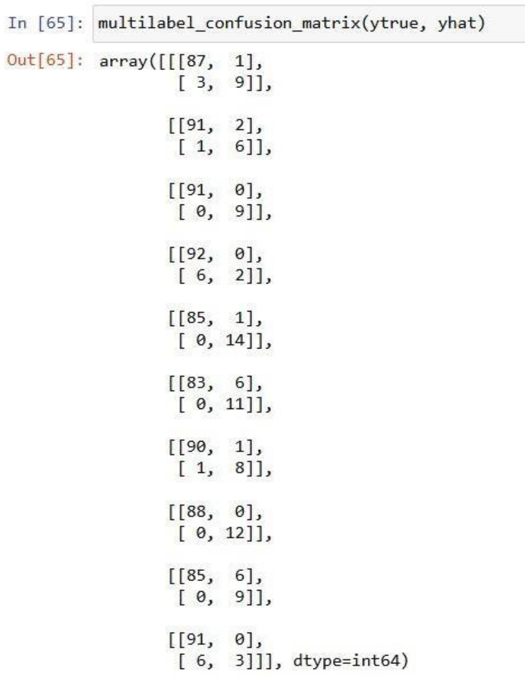
### 5.2 Experimental Result

- Performance on British Sign Language (BSL) Subset:

To test our Long Short-Term Memory (LSTM) model we developed a smaller dataset of 10 unique words from the British Sign Language. We used a confusion matrix to measure the predictions made by the model relative to the actual labels. This matrix gave detailed information on the number of times the model accurately classified each word and showcased the misclassifications.

The confusion matrix showed the strengths of the model to demonstrate legibly specific signs but also areas for improvement. It turned out that certain signs were confused with those having a close semantic meaning due to high similarity in gestures or features, suggesting the need to further increase the dataset size and/or refine structure of model when distinguishing signs with a close meaning.

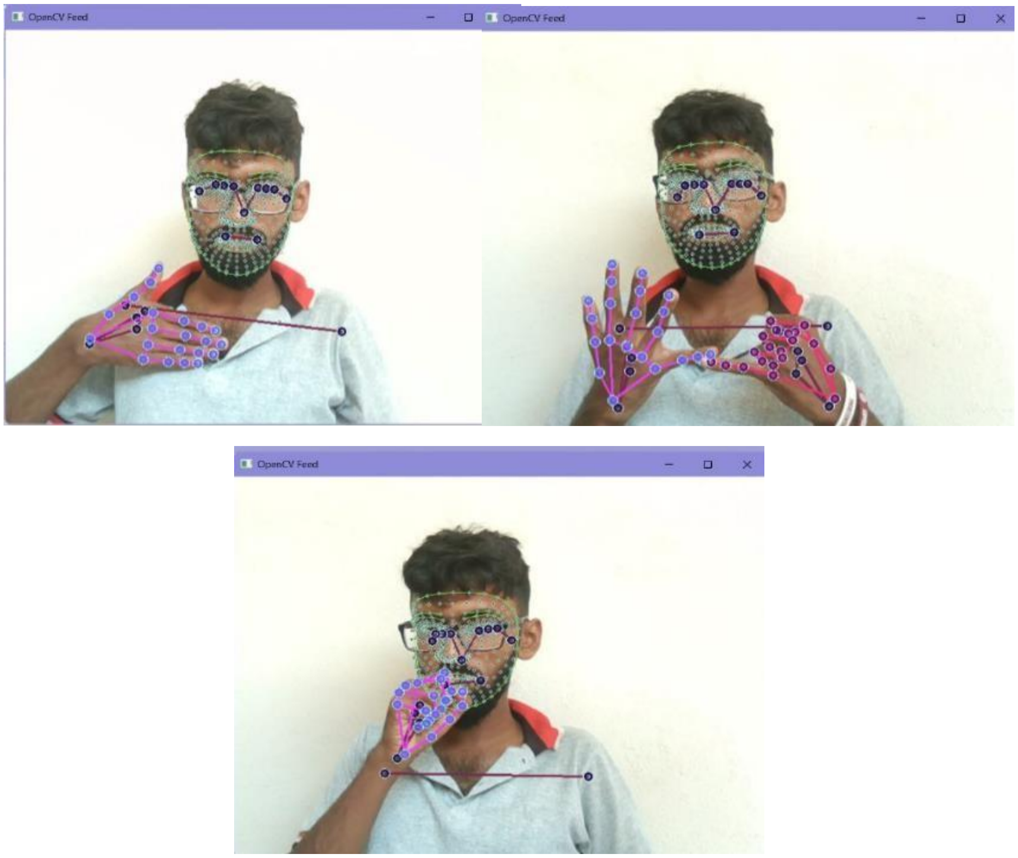
In Figure 16, presents a multilabel confusion matrix-the metric by comparing actual labels to predicted labels on a per-word basis of the whole dataset. Each sub-matrix corresponds to a label and contains four values: true positives, false positives, true negatives, and false negatives. This assessment allows for analyzing the model's capability to predict the correct label when multiple labels exist, highlighting areas where misclassification occurs and improvements can thus be made on prediction accuracy.

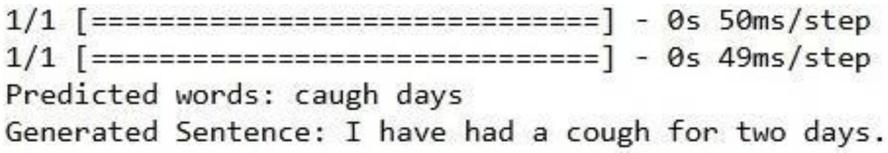


```
In [65]: multilabel_confusion_matrix(ytrue, yhat)

Out[65]: array([[[87,  1],
                 [ 3,  9]],

                [[91,  2],
                 [ 1,  6]],

                [[91,  0],
                 [ 0,  9]],

                [[92,  0],
                 [ 6,  2]],

                [[85,  1],
                 [ 0, 14]],

                [[83,  6],
                 [ 0, 11]],

                [[90,  1],
                 [ 1,  8]],

                [[88,  0],
                 [ 0, 12]],

                [[85,  6],
                 [ 0,  9]],

                [[91,  0],
                 [ 6,  3]]], dtype=int64)
```

**Figure 16 : Visualization of model performance through a confusion matrix, comparing predicted labels to actual labels for each word.**

To demonstrate the practical application of the LSTM model, we implemented a real-time prediction setup using a webcam. The system captures live sign language gestures via the webcam, processes the input through the trained LSTM model, and provides predictions in real time. This setup highlights the model's ability to recognize and classify signs dynamically, showcasing its potential for real-world sign language translation applications.





```
1/1 [==============================] - 0s 50ms/step
1/1 [==============================] - 0s 49ms/step
Predicted words: caugh days
Generated Sentence: I have had a cough for two days.
```

**Figure 17 : Real-Time Sign Language Prediction Using LSTM**

- Performance on HOW2SIGN Dataset :

We delved into two data models-the first based on Transformer architecture and the second on LSTM-aimed at processing sign language video data. For feature extraction, the I3D (Inflated 3D ConvNet) pre-trained model has been used, which has been dubbed one of the best in using videos to infuse temporal-spatial recognition, and features were extracted to be fed into our models.

We trained and evaluated performance on these LSTM and transformers. During the efficacy monitoring of both models, accuracy and loss curves were plotted for training and validation periods. Thus, they help us understand how well these models learn from data and generalization performance on never-before-seen samples.

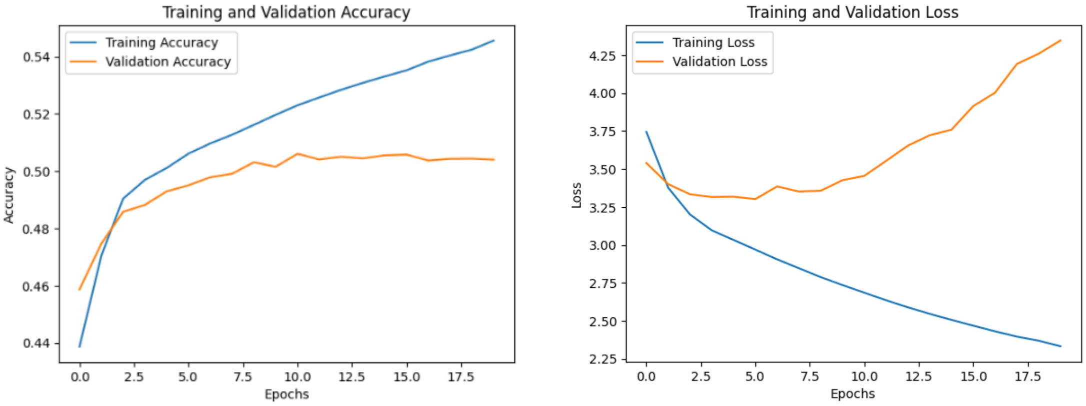


**Figure 18 : Training  and Validation Accuracy Comparison for LSTM**

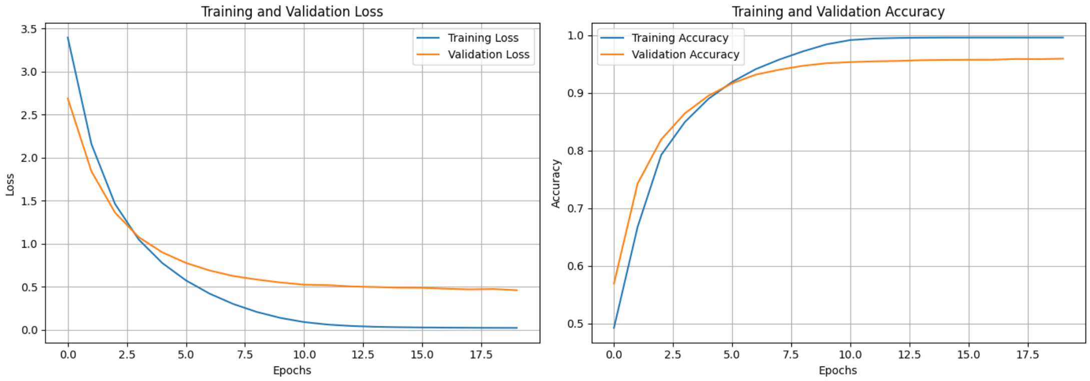


**Figure 19 : Training  and Validation Accuracy Comparison for Transformer**

Additionally, It has also been compared against the I3D feature values from original video data to the feature values taken from our I3D implementation. These features plotted against both sources have drawn a picture of our extraction consistency and reliability.

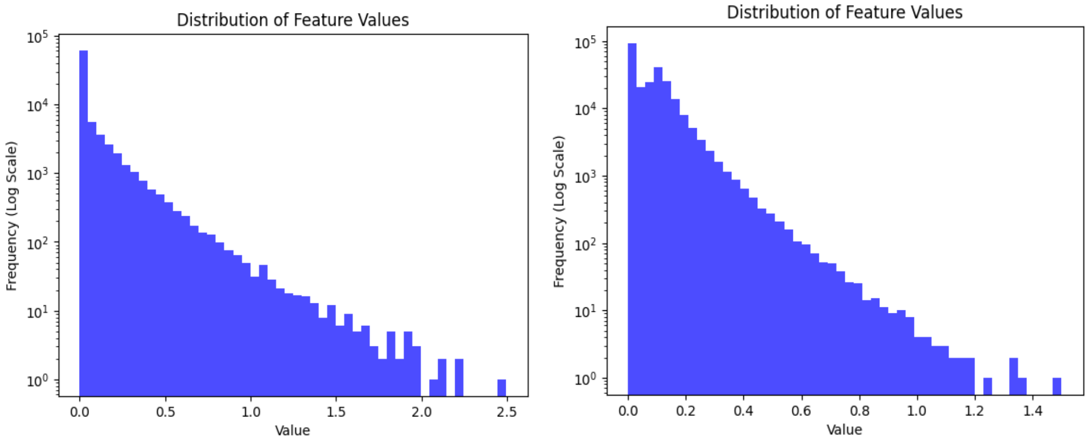


**Figure 20 : For the same video, a comparison of the original I3D features vs. the custom I3D features extracted by our model**

As a next step, we want to enhance the model to do so, we had to train and test 3 models on the HOW2SIGN dataset consisting of 30,384 video samples. The features of I3D extracted from these videos were provided into the models, which performed decoders of different architectures. The evaluation metric utilized for these experiments was the BLEU score, which assesses the quality of generated text in contrast to a benchmark.
The models evaluated were:

1. Encoder + T5 Decoder:

   - Train BLEU Score: 0.0137
   - Validation BLEU Score: 0.0133
   - Test BLEU Score: 0.0129

This model did not fare well, as shown by the low BLEU scores obtained on all the splits. The dependency of the T5 decoder on large sets of training and pre-training conditions might be one of the reasons why it did not perform well on the HOW2SIGN dataset with regards to generalisation.
The T5-based model has a tremendous potential for application in use cases where intensive pre-training and, in addition, larger and more diverse datasets are available.

Nevertheless, it is severely limited in this context by a high dependency on pre-training and a restricted ability to generalize.

2. Encoder + LSTM Layer + Decoder:

    - Train BLEU Score: 0.0388
    - Validation BLEU Score: 0.0299
    - Test BLEU Score: 0.0312

The LSTM layer was added and it did better than the T5 based model in terms of performance. What is striking is that the more text generation there is the lower the BLEU score will be, even though it is low, the text generation is fairly decent. The LSTM layer's capacity to model sequences fits sufficiently well to sign language data's temporal aspects. Here, the architecture establishes the adequateness of LSTM for tasks such as sign language translation, which is primarily temporal modeling. Although this improves the system, there is difficulty encountered with long sequences, and it yielded poor BLEU scores due to having limited training data and intrinsic complexities of precise sign-to-text mapping.

3. Encoder + Transformer Layer + Decoder:

    - Train BLEU Score: 0.0689
    - Validation BLEU Score: 0.0459
    - Test BLEU Score: 0.0478

The Encoder + Transformer Layer + Decoder model demonstrated superior performance compared to the other two models - Encoder + T5 Decoder and Encoder + LSTM Layer + Decoder.

Transformers are highly effective for sequence-to-sequence tasks because of their ability to capture long-range temporal dependencies and contextual relationships between frames in sign sequences. This model efficiently learned both spatial and temporal features from the input sign sequences, resulting in higher BLEU scores and more fluent translations. To ensure scalability and suitability for real-time use, we developed a lightweight version of the system optimized for deployment on mobile and edge devices. The optimization process included reducing model size and improving computational efficiency while maintaining translation accuracy.

A major enhancement during this stage was background removal from each video frame in the dataset. This preprocessing step helped to eliminate unwanted noise caused by irrelevant background elements, which often affected the model's focus and slowed down processing. By isolating only the signer's gestures, the model was able to learn more effectively and operate with faster inference speed during real-time translation.

Figure 21 illustrates an example of an original video frame before preprocessing, where unnecessary background elements could negatively impact model performance. In contrast, Figure 22 shows the corresponding pre-processed frame with the background successfully removed. This visual improvement significantly enhanced gesture visibility, minimized distractions, and improved the overall computational efficiency of the system.

This preprocessing technique was applied to the dataset to enhance the model's real-time operational capacity, making it more robust and adaptable for integration into telemedicine platforms and other real-world healthcare environments. The refined system design demonstrates readiness for scalable deployment in hospitals, clinics, and mobile applications supporting hearing-impaired communication.

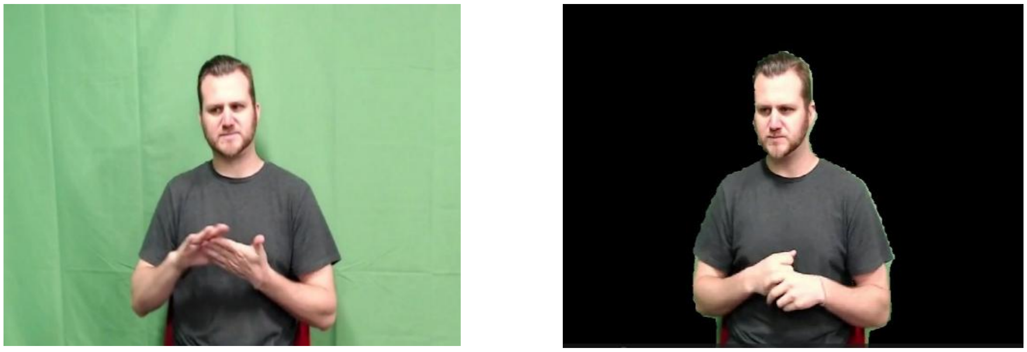


**Figure 21 : Original Video Frame with Background and Processed Video Frame with Background Removed**

Following these preprocessing improvements, the Encoder + Transformer Layer + Decoder model demonstrated a significant boost in translation accuracy.

Here, we present sample predictions generated by the three models using the HOW2SIGN dataset. Each prediction is evaluated against its corresponding reference translation using the BLEU score metric. These examples demonstrate the effectiveness of the models in translating sign language to text and provide insight into their comparative performance.

<u>Encoder + T5 decoder model :</u>

**Table 2 : Sample Predictions and BLEU Scores for Encoder + T5 Decoder on the HOW2SIGN Dataset**

| EXAMPLES | | BLEU |
|---|---|---|
| Reference<br>Prediction | you're only one swing thought away from hooking the **ball** and losing **your** slice **and** this could be it<br>if you have a problem with the **ball** you can do it with the ball or the ball **and you** | 0.05121488961114836 |
| Reference | so the rules get a little bit convoluted but its **important** to remember first of all **to stay on** the strip | 0.048746715608426264 |

| | | |
|---|---|---|
| Prediction | I mean it is **important** that you stay on top of that and you have **to stay on** top of that so you | |
| Reference<br>Prediction | but what it'll do is it'll bring up a window on your **computer** that brings up the task manager<br>I am going to do a **computer** on my computer | 0.01213507123352864 |
| Reference<br>Prediction | don't worry about your **power** don't worry about getting everything perfect just snap off **a lot of** punches<br>I want to give you a little bit of clarity on how to play **a lot of power** and power | 0.05121488961114836 |
| Reference<br>Prediction | and **doing** this **could be a little harder in** smaller **vehicle** but it still could be done<br>a lot of different things that you can **do in a car can be difficult** | 0.01522433772663941 |
| Reference<br>Prediction | but **a good** player that has the strokes once they get the wheel chair down **you're** in trouble<br>**a great** idea to get **your** hair out of the way is to have a squeegee | 0.014286401328894116 |
| Reference<br>Prediction | **this is a flamingo catch**<br>**a little bit more relaxed** | 0.05372849659117709 |

Encoder + Transformer layer + decoder model :

**Table 3 : Sample Predictions and BLEU Scores for Encoder + Transformer Layer + Decoder on the HOW2SIGN Dataset**

| EXAMPLES | | BLEU |
|---|---|---|
| Reference | **you're** only one swing thought **away** from hooking the **ball and** losing your slice and this could be it | 0.0601429426464788 |

| | | |
|---|---|---|
| Prediction | if **you're** not going to hit the **ball and** you want to be able to take it **out** | |
| Reference<br>Prediction | so the rules get a little bit convoluted but its **important to remember** first of all **to stay on the strip**<br>I think its **important to remember to keep** the thread on the other side so that you can **stay on the other side** | 0.1629944673128894 |
| Reference<br>Prediction | but what it'll do is it'll bring up **a** window on **your computer** that brings up the task manager<br>I have **a** process that you can do on **your computer** | 0.05361218207146106 |
| Reference<br>Prediction | don't worry **about** your **power** don't worry **about** getting everything perfect just snap off a lot **of** punches<br>I am going to talk **about** the **power** swings and the balance **of** the chest | 0.014242474285751547 |
| Reference<br>Prediction | and doing this could be **a little** harder **in smaller vehicle** but it still could be done<br>a lot of people are going to need a **car** that has **a little bit of** a little bit of | 0.020364851292391 |
| Reference<br>Prediction | but a good player that has the strokes once they get the wheel chair down, you're in trouble<br>I have a squat so I have a squat so I have | 0.01033114956441737 |

Encoder + LSTM layer + decoder model :

**Table 4 : Sample Predictions and BLEU Scores for Encoder + LSTM Layer + Decoder on the HOW2SIGN Dataset**

| EXAMPLES | | BLEU |
|---|---|---|
| Reference<br>Prediction | **you're** only one **swing** thought away from hooking **the ball and** losing your slice and this **could be it**<br>if you want to hit **the ball** you want to take **your swing** to the right side of the ball so **you can** | 0.020828838183973034 |
| Reference<br>Prediction | so, the rules get a little bit convoluted but it's **important to remember** first of all **to stay on the strip**<br>a lot of times it is very **important to keep in mind** that it is important to keep in mind that it is | 0.022023814946586635 |
| Reference<br>Prediction | but what it'll do is it'll bring up a window on your **computer** that brings up the task manager<br>I have a **computer** that has a computer that has a computer that has a computer that has | 0.023914960914330066 |
| Reference<br>Prediction | don't worry about your power don't worry about getting everything perfect just snap off a lot of punches<br>I really want to make sure that you have a good workout on the back of the bike | 0.01284618972676772 |
| Reference<br>Prediction | and doing this could be **a little** harder **in smaller vehicle** but it still could be done<br>if you're going to be **a little** older you might be able to get **a bigger car** and you | 0.057259987315337754 |

We compared our BLEU scores with previously reported state-of-the-art (SOTA) results. Prior studies on similar sentence-level sign language translation tasks have reported maximum BLEU scores up to 8.03.

**Table 5 : Comparison of BLEU Scores with State-of-the-Art (SOTA)**

| Model | Train | Validation | Test |
|---|---|---|---|
| Encoder + T5 Decoder | 0.0137 | 0.0133 | 0.0129 |
| Encoder + LSTM Layer + Decoder | 0.0388 | 0.0299 | 0.0312 |
| Encoder + Transformer Layer + Decoder | 0.0689 | 0.0459 | 0.0478 |
| SOTA (Reference from Previous Research) | - | - | 0.0803 |

# Chapter 6: CONCLUSIONS AND FUTURE DIRECTIONS

## 6.1 Summary and Conclusion

The outcome of this research highlights the critical influence of model architecture in the field of sign language translation using video datasets. Three architectures were evaluated:

- Encoder + T5 Decoder
- Encoder + LSTM Layer + Decoder
- Encoder + Transformer Layer + Decoder

Among these, the Transformer-based model demonstrated superior performance across training, validation, and testing stages. This improvement is primarily due to the temporal dependencies that Transformer architectures effectively model, allowing them to handle sequential data such as sign language videos more efficiently than traditional LSTM-based or T5 decoder approaches.

The Transformer model's ability to capture hand gesture progressions, facial expressions, and contextual motion patterns contributed to achieving more accurate and meaningful sentence-level translations. However, despite the performance improvement, the BLEU scores remained relatively low for all models. This can be attributed to several key factors:

Limitations of the Dataset: Although the HOW2SIGN dataset is extensive, it lacks diversity in signing styles, signer backgrounds, and dialect variations, which limits model generalization.

Translation Complexity: Translating sign language gestures into grammatically correct natural language sentences involves intricate linguistic structures and spatial-temporal understanding, making the task inherently challenging.

Model Optimization: The models may be underfitted due to suboptimal hyperparameter selection. Further optimization or hybrid architectures could enhance translation fluency and accuracy.

In addition to model evaluation, the project successfully developed a real-time two-way desktop communication system using socket communication, integrating the Transformer-based model for real-time translation between patients and doctors.
While the system functioned effectively, latency issues were observed during testing—mainly due to the computationally intensive nature of Transformer-based video processing. Although the delay was not critical, it highlighted the importance of further optimization for low-latency performance to ensure smooth real-time interaction in telehealth settings.

Overall, this project achieved its key objectives by implementing and evaluating deep learning models capable of continuous sign language translation. The system demonstrated the feasibility of integrating AI-driven translation models into assistive telehealth applications that can significantly improve communication accessibility for individuals with hearing and speech impairments.

## 6.2 Future Work

To further enhance the performance, scalability, and accessibility of the proposed real-time sign language translation system, several directions are identified for future development:

1. Dataset Expansion

Future work will focus on expanding the dataset to include more diverse and localized data, particularly Sinhala Sign Language (SSL) videos. This enhancement will allow the model to better generalize to real-world scenarios and improve accuracy across different signers, environments, and dialects. By incorporating local sign language datasets, the system can become more inclusive and culturally relevant for Sri Lanka's healthcare context.

2. Integration with Cloud and Mobile Platforms

The current solution is designed as a desktop-based application for telehealth communication. As part of the future development, the system will be extended to cloud-based and mobile platforms to improve scalability and accessibility. A cloud-integrated solution will allow real-time translation and communication over remote servers, while a mobile version will ensure convenience and portability for users, enabling them to use the system from any location.

3. Enhanced Bi-Directional Communication with Sign Language Avatar

Currently, the two-way desktop communication system allows the patient to perform sign gestures in front of a camera, which are translated into text and displayed on the doctor's desktop interface. The doctor can then respond using voice messages or quick text replies, which are displayed back on the patient's side.

As a future enhancement, the system will incorporate a sign language avatar model to convert the doctor's spoken or text responses into animated sign gestures. This improvement will allow deaf patients to see and understand the doctor's messages through visual sign representation rather than reading text. Such an addition would make the communication process faster, more natural, and easier to comprehend for users who primarily rely on sign language.

# REFERENCES

[1] M. S. Astriani and M. A. R. C. E. L. L. Alvianto, "Telemedicine sign language classification for COVID-19 patients with disability based on LSTM model.," 2023.

[2] K. Yin and J. Read, "Attention is all you sign: sign language translation with transformers.," *Sign Language Recognition, Translation and Production (SLRTP) Workshop-Extended Abstracts.,* vol. Vol. 4, 2020, August.

[3] O. Sincan, N. Camgoz and R. Bowden, "Using an LLM to Turn Sign Spottings into Spoken Language Sentences.," *arXiv preprint arXiv:2403.10434,* 2024 .

[4] N. Aloysius , M. Geetha and P. Nedungadi , "Incorporating relative position information in transformer-based sign language recognition and translation.," *IEEE Access,* vol. 9, pp. 145929-145942, 2021 Oct 26.

[5] S. Wazalwar and U. Shrawankar , "Online healthcare consultation system for deaf & dumb during pandemic situation," *Bioscience Biotechnology Research Communications,* vol. 13, no. 14, pp. 213-216, 2020.

[6] Y. Zhao , . X. Zhang , R. Hu , J. Xue , X. Li , . L. Che , R. Hu and . L. Schopp , "An automatic captioning system for telemedicine.," *In 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings,* vol. vol. 1, pp. pp. I-I, 2006.

[7] . M. De Coster , . M. Van Herreweghe and J. Dambre, "Sign language recognition with transformer networks.," *12th international conference on language resources and evaluation. European Language Resources Association (ELRA),* pp. pp. 6018-6024, 2020.

[8] . H. Nagendraswamy and B. Kumara , "LBPV for recognition of sign language at sentence level: An approach based on symbolic representation.," *Journal of Intelligent Systems,* pp. pp.371-385, 2017.

[9] A. Moryossef , . I. Tsochantaridis , J. Dinn , . N. Camgoz , R. Bowden , . T. Jiang , A. Rios , . M. Muller and . S. Ebling , "Evaluating the immediate applicability of pose estimation for sign language recognition.," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.,* pp. pp. 3434-3440, 2021.

[10] S. Ko , . C. Kim , . H. Jung and C. Cho , "Neural Sign Language Translation Based on Human Keypoint Estimation," *Applied sciences 9.13 (2019),* p. p.2683, 2019.

[11] . L. Tarrés , G. Gállego , A. Duarte , . J. Torres and X. Giró-i-Nieto , "Sign language translation from instructional videos.," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.,* pp. pp. 5625-5635, 2023.

[12] G. Tanzer , M. Shengelia , K. Harrenstien and D. Uthus , "Reconsidering Sentence-Level Sign Language Translation.," *arXiv preprint arXiv:2406.11049,* 2024.

[13] P. Alvarez , X. Giro , N. Laia and T. Benet , "Sign Language Translation based on Transformers for the How2Sign Dataset.," *Image Processing Group Signal Theory and Communications Department Universitat Politècnica de Catalunya. BARCELONATECH,* 2022.

[14] P. Fayyazsanavi , A. Anastasopoulos and J. Košecká , "Gloss2Text: Sign Language Gloss translation using LLMs and Semantically Aware Label Smoothing.," *arXiv preprint arXiv:2407.01394,* 2024.

[15] " Universitas Sebelas Maret," UNS Students Successfully Develop Sign Language Translator Gloves, 2020. [Online]. Available: https://uns.ac.id/en/uns-students-successfully-develop-sign-language-translator-gloves/.

[16] Z. Zhou , K. Chen , X. Li , S. Zhang , Y. Wu , Y. Zhou , K. Meng , C. Sun , . Q. He, W. Fan and E. Fan , "Sign-to-speech translation using machine-learning-assisted stretchable sensor arrays.," *Nature Electronics,* 2020.

[17] I. Godage , "Sign Language Recognition for Sentence Level Continuous Signings," *Doctoral dissertation,* 2021.

[18] . R. Rishan , . S. Jayalal and . T. Wijayasiriwardhane , "Translation of sri lankan sign language to sinhala text: A leap motion technology-based approach.," *In 2022 2nd International Conference on Advanced Research in Computing (ICARC),* pp. pp. 218-223, 2022.

[19] T. Chong and B. Lee , "American sign language recognition using leap motion controller with machine learning approach.," *Sensors,* p. p.3554, 2018.

[20] J. Bird , A. Ekárt and D. Faria , "British sign language recognition via late fusion of computer vision and leap motion with transfer learning to american sign language.," *Sensors 20, no. 18,* p. p.5151, 2020.

[21] "BOBSL (BBC-Oxford British Sign Language)," paperswithcode.com, [Online]. Available: https://paperswithcode.com/dataset/bobsl.

[22] *RWTH-PHOENIX-Weather 2014*. (n.d.). Retrieved from https://paperswithcode.com/dataset/rwth-phoenix-weather-2014

[23] *How2Sign Dataset*. (n.d.). Retrieved from https://how2sign.github.io/

[24] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," CoRR, vol. abs/1910.13461, Oct. 2019. [Online]. Available: http://arxiv.org/abs/1910.1346