Deep Learning-Based Real-Time Sign Language Translation System for Telehealth Applications

Gunarathna L.P.N, Somarathna S.V.A.P.K, T. Mukunthan
Department of Electrical and Electronic Engineering
University of Jaffna, Sri Lanka
2020e046@eng.jfn.ac.lk, 2020e212@eng.jfn.ac.lk, mukunthan@eng.jfn.ac.lk

***Abstract*—Every human being requires communication; however, people with hearing and speech disabilities have difficulty conveying messages, particularly in urgent and professional settings like the health sector. In this study, we propose a deep learning-based real-time sign language translation system specifically designed for Sri Lanka's telehealth applications. The system extracts spatial-temporal features with an I3D (Inflated 3D ConvNet) and tests three sequence-to-sequence models: T5, LSTM, and Transformer decoders. Out of the three evaluated models in the paper, the Encoder-Transformer-Decoder performed the best relative to the other models evaluated. To allow instant interaction for both patients and professionals, a real-time bi-directional desktop application using socket communication was implemented. The system's main focus is prompt interaction with an emphasis on Sri Lanka's telehealth translation demands, providing high-speed translation, accessibility, and accuracy at the sentence level. The work demonstrates some promise, but the remaining issues of the lack of sufficient datasets and the need for localization of the language still pose problems. As a result, integrating multiple languages and modalities and improving the real-world usefulness and the quality of the translation of the system remain the focus of future work.**

***Index Terms*—Sign Language Translation (SLT), Deep Learning, Human-Computer Interaction, Transformer, LSTM, Telehealth Applications, Assistive Technology**

I. Introduction

Communication essentially matters for social inclusion [1], and hearing- and speech-impaired individuals face specific hurdles in vital aspects of life, such as healthcare. While there exist traditional, hardware-based systems for sign language translation, these usually suffer from lack of real-time performance, lack of usability, and lack of sentence-level comprehension. However, the research in Sri Lanka appears to be restricted to basic gesture recognition with limited vocabulary and accuracy.

Therefore, this system proposes a deep learning-based real-time sign language translation system for the healthcare-related applications. Using the HOW2SIGN dataset [2] with I3D features [3], three architectures were trained and evaluated: T5, LSTM, and Transformer-based decoders. Consequently, the Encoder + Transformer gave the best performance among all tested models. Using socket-based communication, real-time translation takes place between patients and doctors in support of bi-directional communication. Therefore, some of the limitations of current solutions are addressed by looking into sentence-level accuracy, practical usability, and real-time response.

The rest of this article is organized as follows. Section II presents the related works in sign language translation. The detailed methodology is presented in Section III. Experimental results are presented in Section IV, followed by conclusions and future work in Sections V and VI respectively.

II. Related Works

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 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 new directions in 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.

A. Context-Aware Sign Language Translation

To investigate if the discourse level influences the translation of sign language, researchers from Google collaborated with Rochester Institute of Technology [15]. Garrett Tanzer, Maximus Shengelia, Ken Harrenstien, and David Uthus offered a new human reference for ASL to English translation as a result of the How2Sign dataset. In this work, they looked at the long-range syntactic relation, thus, unlike previous work, not concerned with sentence clipping. Their approach was to employ interpreters who are proficient in ASL-English to translate ASL clips without concerning the rest of the document.

As a result, they concluded that information from 33% of the sentences is needed for further clarification. 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. This paper stresses the significance of the context in translating sign language and also the difficulties in such process for both human and machine.

B. Transformer-Based Approaches

Patricia Cabot Alvarez, Xavier Giro Nieto, and Laia Tarres Benet proposed a Transformer-based approach for the

How2Sign dataset [16]. First of all, they tested 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 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 proved useful in providing foundation for subsequent researches in sign language translation.

C. Real-Time Classification Systems

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 Bina Nusantara University, Indonesia using BISINDO data [17]. Their methodology involves several key steps: installation of OpenCV and MediaPipe Holistics for data preprocessing and gathering, use of LSTM together with multiple Dense layers for training.

Therefore, the systematic detection provided for sign languages of speech and hearing-impaired individuals established a more effective approach during COVID-19. The performance for training data set, testing data set, and real-time classification achieved 99%, 98%, and 90% accuracy respectively. The high performance established that the sign language using the LSTM architecture distinguishes the signs and even terms referring to COVID-19.

D. Research Gaps Identified

From the reviewed literature, several gaps remain:

- **Lack of localized datasets** – Most existing works rely on ASL-based datasets, limiting the applicability to regional sign languages such as Sri Lankan Sign Language (SLSL).
- **Limited sentence-level translation** – Prior studies focused mainly on isolated sign recognition rather than full sentence translation required in real-world communication.
- **Insufficient multimodal integration** – Few studies combine skeleton, optical flow, and spatial-temporal cues to capture the rich dynamics of sign gestures.
- **Lack of real-time telehealth implementation** – Existing systems rarely evaluate latency, usability, or performance in healthcare environments.
- **Weak contextual learning** – Many previous models overlook discourse-level context and fail to integrate large language models for enhanced semantic understanding.

E. Our Contribution and Novelty

To address these limitations, our study proposes a deep learning-based real-time sign language translation system for telehealth applications. Unlike previous works, our system:

- Utilizes I3D feature extraction combined with optical flow and skeletal keypoints for robust motion representation
- Incorporates a large language model (LLM) to enhance translation fluency and contextual accuracy
- Evaluates multiple architectures (T5, LSTM, Transformer) for comparative performance
- Implements real-time two-way translation using socket-based communication
- Explores telehealth-specific latency and deployment feasibility

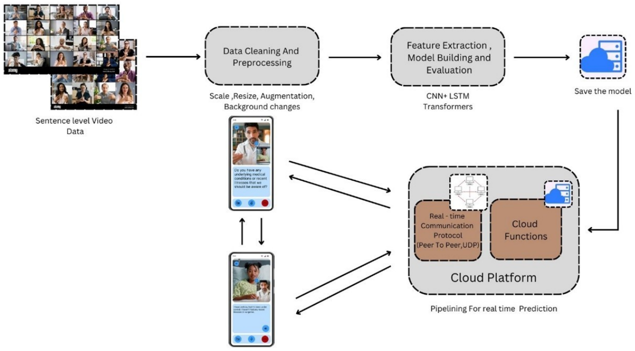
In this way, this approach bridges the gap between research-oriented sign translation models and practical healthcare applications, moving toward inclusive, real-time doctor-patient communication systems.

III. Methodology

This section will explain the details of the sign language translation system, including the overall system architecture, data acquisition and preprocessing techniques, I3D-based feature extraction pipeline, sequence modeling using deep learning architectures, performance evaluation metrics, and real-time mobile application deployment strategy.

A. System Overview

The proposed system aims to provide real-time sign language translation in the medical field, improving communication quality between healthcare staff and hearing-impaired patients. Therefore, the system architecture is composed of five main stages: data acquisition of sign language, preprocessing and data augmentation, feature extraction through an Inflated 3D ConvNet (I3D), sequence modeling by deep neural networks, and real-time translation implementation on mobile devices.



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?
Real - time Communication Protocol (Peer To Peer,UDP)
Cloud Functions
Cloud Platform
Pipelining For real time Prediction

Fig. 1. Overall System Architecture

As illustrated in Fig. 1, video input of sign language is captured using a mobile device and processed by a multi-stage pipeline. Following spatial-temporal feature extraction, complex sequence models convert gestures into English text and show them on the doctor's display. They can further be synthesized into audible speech if needed.

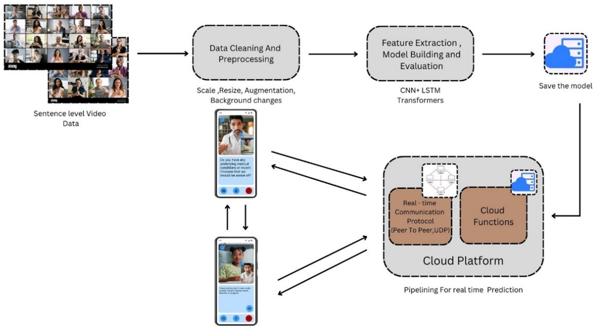

Sentence level Video Data
Data Cleaning And Preprocessing
Scale ,Resize, Augmentation, Background changes
Feature Extraction , Model Building and Evaluation
CNN+ LSTM Transformers
Save the model
Do you have any underlying medical conditions or recent illnesses that we should be aware of?
Real - time Communication Protocol (Peer To Peer,UDP)
Cloud Functions
Cloud Platform
Pipelining For real time Prediction

Fig. 2. System Overview

B. Data Acquisition and Preprocessing

1) Dataset Selection: Three publicly available datasets were analyzed for suitability:

- **BBC-Oxford BSL Dataset** [4]: Rich in isolated and continuous sign sequences with subtitle alignment.
- **RWTH-PHOENIX-Weather 2014T Dataset** [5]: Annotated German sign language corpus with glosses and non-manual features.
- **How2Sign Dataset** [2]: A large-scale, multimodal American Sign Language (ASL) dataset containing over 80 hours of RGB and depth video, aligned with English translations.

Consequently, the How2Sign dataset was chosen because it covers sentence-level sign sequences comprehensively and provides multimodal data including depth and pose.

2) Preprocessing Pipeline: Every video goes through an extensive preprocessing process:

- **Background Removal:** Implementing OpenCV to segment signers from intricate backgrounds.
- **Skeleton Extraction:** MediaPipe Holistic modules are used to detect human body, hand, and face keypoints, which are then converted to structured skeletal data.

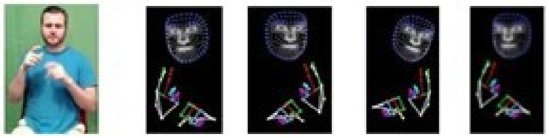


Fig. 3. Skeleton Extraction and Augmentation

- **Optical Flow Calculation:** Frame-wise motion vectors are calculated based on the TV-L1 [6] algorithm. They depict the displacement and direction of the hand movements.

Mathematically, optical flow is modeled by the Optical Flow Constraint Equation:

$$I_x \cdot \frac{dx}{dt} + I_y \cdot \frac{dy}{dt} + I_t = 0 \tag{1}$$

where I_x and I_y represent spatial gradients, and I_t represents the temporal gradient.

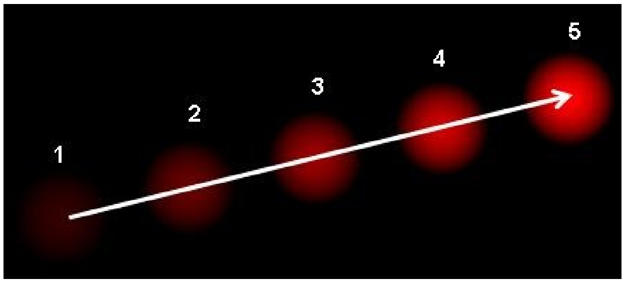


Fig. 4. Optical Flow Vector Analysis

C. Feature Extraction using I3D

We use the Inflated 3D ConvNet (I3D) model for spatial-temporal encoding of gesture motion. As a result, I3D inflates pre-trained 2D CNN filters to 3D to learn from video volumes, effectively capturing both spatial appearance and temporal motion patterns simultaneously.

D. Sequence Modeling

We explore multiple deep sequence architectures to map I3D feature sequences to English translations:

1) GRU Model: Gated Recurrent Units (GRUs) [7] are computationally efficient and suitable for temporal sequences of short and medium lengths. The model captures dependencies by updating hidden states at every process step according to the sequence of input features.

2) LSTM Model: Long Short-Term Memory (LSTM) networks [8] improve upon GRU by introducing memory cells and gating functions (input, forget, output), efficiently dealing with long-term dependencies and vanishing gradients.

3) Transformer Model: Transformers [9] employ self-attention to model global context over the whole sequence. In this way, every frame attends to every other frame, making it easy to deal with gestures of variable length. Temporal ordering is preserved using positional encodings.

4) Hybrid Model (LSTM + Transformer): A hybrid approach in which LSTM captures local temporal features and a Transformer decoder performs sequence-to-sequence translation. Model training takes I3D features as input and English sentences as output.

E. Evaluation Metrics

The following parameters are taken to quantitatively analyze the performance and real-world applicability of the sign language translation model:

1) **BLEU Score** [10]: The Bilingual Evaluation Understudy (BLEU) score is crucial in evaluating machine translation systems by assessing n-gram overlap between predicted outputs and reference sentences.
2) **Word Error Rate (WER)** [11]: WER computes the total number of substitutions (S), deletions (D), and insertions (I) required to match the predicted sentence to the reference, normalized by the number of words (N) in the reference:

$$WER = \frac{S + D + I}{N} \tag{2}$$

3) **Sentence Error Rate (SER)** : The Sentence Error Rate measures how many sentences contain at least one error. A lower SER reflects better overall sentence correctness and contextual reliability.
4) **Latency:** Latency describes the period of time taken to capture video input and the subsequent processing of the final text output. This is crucial in time-sensitive channels like medicine.

F. Mobile Deployment Strategy

The completed model is integrated into a mobile application using an appropriate framework. Therefore, the application allows for real-time bidirectional communication between doctor and patient. The key design features are:

- **Edge or Cloud Inference:** The model can be executed on-device or utilize cloud capabilities based on hardware limitations.
- **Peer-to-Peer Communication:** WebRTC is used to send real-time video and translation data.
- **UI Design:** The top half is taken by the video, and the translated text is placed in the lower half. Multimodal interaction is facilitated through speech-to-text and text-to-speech modules.

In this way, the system ensures low-latency translation and is optimized for use in hospital and telehealth scenarios.

IV. Experiments

This section outlines the development of a sign language translation system, starting with word-level recognition using an LSTM model and highlighting its limitations. To address this, sentence-level translation was implemented using the HOW2SIGN dataset and evaluated across T5, LSTM, and Transformer models using BLEU scores. As a result, a real-time two-way communication system was also developed to demonstrate practical use in telehealth settings.

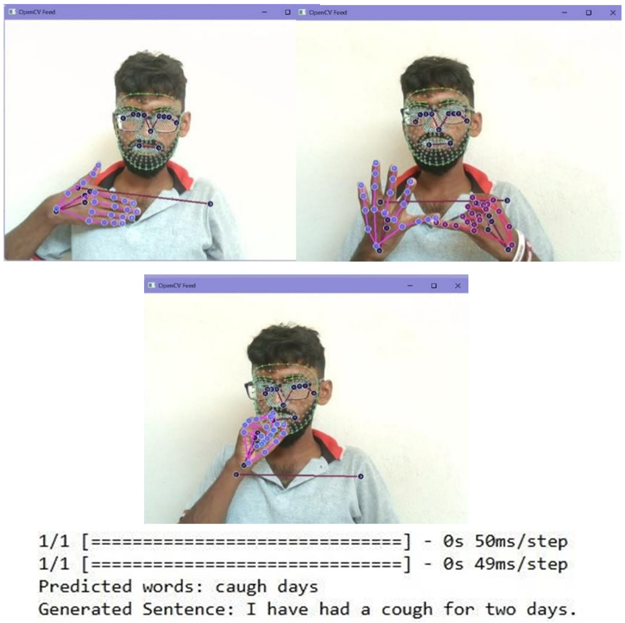
A. Initial Word-Level Experiments

In the beginning, we developed a word-level sign language dataset of 10 common BSL signs relating to telehealth with 2,000 video samples. Skeletal landmarks were extracted using MediaPipe. These landmarks were fed to an LSTM over 50 epochs using categorical cross-entropy loss with Adam optimizer. Consequently, the model predicted individual signs well in real-time but could not translate them into meaningful sentences.

To overcome this limitation, a pretrained Large Language Model (LLM) [12], BART, was fine-tuned with example telehealth sentence structures. LSTM model output word sequences were inputted to the LLM to generate meaningful and grammatically correct sentences.

Yet, the transition exposed typical limitations exhibited in word-level data, including:

- Poor contextual understanding
- Inappropriate grammar
- Too much dependence on advanced language models for producing coherent sentences



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

Fig. 5. Real-Time Sign Prediction Using Webcam and LSTM Model

B. Transition to Sentence-Level Dataset

Understanding the limitations of word-level methods, we moved to a sentence-level dataset HOW2SIGN to allow richer context comprehension and translation on a full-sentence level.

Key advantages of HOW2SIGN:

- Direct Accessibility: Publicly available, saving time on data collection
- Rich Multimodal Data: Comprised of videos, audio, and transcriptions in English
- Sentence-Level Annotations: Supports grammatically and contextually accurate sentence generation
- 3D Pose Estimation: Over 3 hours of skeletal captures aid in spatio-temporal gesture understanding
- Diverse Vocabulary: More than 33,000 data entries spanning 24,703 distinctive words across diverse contexts

C. Model Training and Evaluation

We utilized the HOW2SIGN dataset which contains 30,384 samples of sign language performing videos. Once we extracted spatial-temporal features with the I3D approach, we trained and assessed three different models in the sequence-to-sequence paradigm:

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

All models were trained and tested using the standard dataset splits. Model training was performed on Google Colab utilizing a GPU-enabled environment (NVIDIA Tesla T4, 16 GB VRAM) with 12 GB of system RAM, which ensured efficient handling of the large-scale video data and reduced training time.

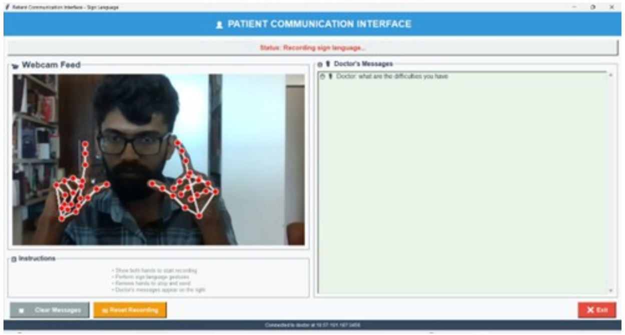
The performance of the models was evaluated using the BLEU score, a standard metric for assessing the quality of machine translation outputs. Among the three architectures, the Encoder + Transformer + Decoder model demonstrated superior performance by effectively capturing temporal sign dynamics and producing the most accurate translations.

D. *Real-Time Implementation*

To validate the real-world applicability of our trained model, we developed a real-time two-way communication system using socket programming. This setup simulates a remote healthcare scenario and consists of two components: a patient-side module and a doctor-side module.

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
Status: Recording sign language...
Webcam Feed
Doctor's Messages
Instructions
Clear Messages
Reset Recording
✕ Exit

Doctor: what are the difficulties you have

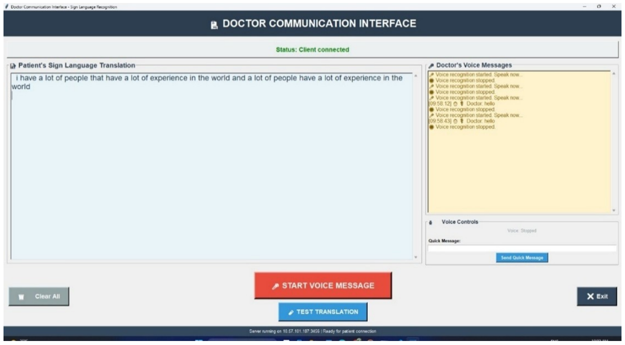
- Show both hands to start recording
- Perform sign language gestures
- Remove hands to stop and send
- Doctor's messages appear on the right

Connected to doctor at 10.57.101.187:3456

Fig. 6. Patient-Side Sign Language Video Capture and Transmission

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



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.
```

Server running on 10.57.101.187:3456 | Ready for patient connection

Fig. 7. Doctor-Side Sign Language Video Processing and Translation

V. Results

A. *Performance on British Sign Language (BSL) Subset*

For testing the first word-level LSTM model, we compiled a dataset containing 10 telehealth-related British Sign Language [13] vocabulary words frequently used in the British telehealth context. Its classification performance was evaluated using a confusion matrix, which reveals both hits and misses.

The LSTM-based model's performance was satisfactory for most distinct signs; however, it was poor at classifying ”Me” and ”My”. A multilabel confusion matrix [14] provided an in-depth analysis and exposed areas of sign prediction accuracy. The prototype's ability to perform sign word recognition and display words dynamically through webcam input was a significant milestone.

B. *Performance on HOW2SIGN Dataset*

We used the HOW2SIGN dataset, which consists of over 30,000 video samples, to evaluate sentence-level sign language translation. Initially, we checked the integrity of our I3D-based feature extraction by comparing feature vectors from our model with product features vectors in the original dataset, which proved our feature extraction pipeline is reliable and valid.

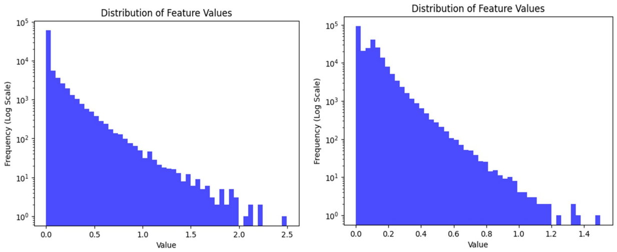


Fig. 8. Comparison of original I3D features vs. custom I3D features

We trained and evaluated two basic sequence models: LSTM-based and Transformer-based architectures. Both models used the pretrained I3D network to extract temporal-spatial features. Accuracy and loss curves during training revealed that the LSTM better captured temporal dependencies, while the Transformer handled longer sequences more effectively.

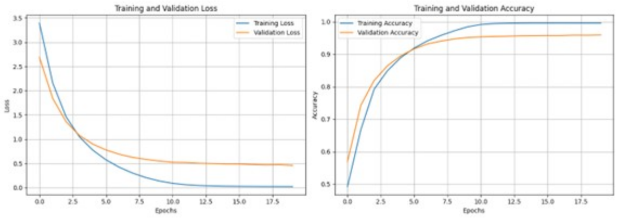


Fig. 9. Training and Validation Accuracy Comparison for Transformer

For refining precision and improving the flow of sentence-level sign language translation, we created and experimented with three models that differed only in decoder structure. Model performance was evaluated with BLEU scores.

The T5-based model was at the lowest point, probably because it depended heavily on pre-training and huge, varied datasets which were not the case in the HOW2SIGN data. Therefore, the Encoder + Transformer + Decoder model

TABLE I
BLEU SCORE COMPARISON OF MODEL ARCHITECTURES

Model	Train	Val	Test
Encoder + T5 Decoder	0.0137	0.0133	0.0129
Encoder + LSTM + Decoder	0.0388	0.0299	0.0312
Encoder + Transformer + Decoder	**0.0689**	**0.0459**	**0.0478**

recorded the highest BLEU scores and performed more effectively across the different dataset splits, thus capable of representing the temporal structure of the sign language.

C. Comparison with State-of-the-Art

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 II
COMPARISON WITH STATE-OF-THE-ART

Model	Train	Val	Test
Our: Encoder + T5	0.0137	0.0133	0.0129
Our: Encoder + LSTM	0.0388	0.0299	0.0312
Our: Encoder + Transformer	0.0689	0.0459	0.0478
SOTA (Previous Research)	-	-	0.0803

VI. CONCLUSION

This research evaluated three architectures for sign language translation: Encoder + T5 Decoder, Encoder + LSTM Layer + Decoder, and Encoder + Transformer Layer + Decoder. The Transformer-based model achieved superior performance by effectively capturing temporal dependencies in sign language videos, including hand gestures, facial expressions, and motion patterns.

Despite this improvement, BLEU scores remained relatively low due to dataset limitations, translation complexity, and potential model underfitting. The project successfully developed a real-time two-way communication system using socket communication and the Transformer model. While functional, latency issues highlighted the need for optimization in computationally intensive video processing.

Overall, this project demonstrates the feasibility of AI-driven sign language translation in assistive telehealth applications, improving communication accessibility for individuals with hearing and speech impairments.

VII. 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:

A. Dataset Expansion

Future work will focus on expanding the dataset to include more diverse and localized data, particularly Sinhala Sign Language (SSL) videos. Therefore, 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.

B. Enhanced Bi-Directional Communication

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. In this way, 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] "UCSF Health," University of California San Francisco. [Online]. Available: https://www.ucsfhealth.org/education/communicating-with-people-with-hearing-loss

[2] "How2Sign Dataset." [Online]. Available: https://how2sign.github.io/

[3] "Video Features Documentation - I3D." [Online]. Available: https://viashin.github.io/video_features/models/i3d/

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

[5] "RWTH-PHOENIX-Weather 2014 Dataset." [Online]. Available: https://paperswithcode.com/dataset/rwth-phoenix-weather-2014

[6] A. Wedel, T. Pock, C. Zach, H. Bischof, and D. Cremers, "An improved algorithm for tv-l1 optical flow," in *Statistical and Geometrical Approaches to Visual Motion Analysis*, Springer Berlin Heidelberg, 2009, pp. 23–45.

[7] S. Chakraborty, P. Paul, S. Bhattacharjee, S. Sarkar, and A. Chakraborty, "Sign Language Recognition Using Landmark Detection, GRU and LSTM," *American Journal of Electronics & Communication*, pp. 20–26, 2023.

[8] T. Liu, W. Zhou, and H. Li, "Sign language recognition with long short-term memory," in *2016 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2016, pp. 2871–2875.

[9] N. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10023–10033.

[10] "Deepchecks - BLEU Score." [Online]. Available: https://www.deepchecks.com/glossary/bleu/

[11] "Deepchecks - Word Error Rate (WER) Score." [Online]. Available: https://www.deepchecks.com/glossary/word-error-rate-wer-score/

[12] "What is a Large Language Model?" AWS. [Online]. Available: https://aws.amazon.com/what-is/large-language-model/

[13] "British Sign Language Dictionary," SignBSL.com. [Online]. Available: https://www.signbsl.com/

[14] "Understanding the Confusion Matrix in Machine Learning." [Online]. Available: https://www.geeksforgeeks.org/machine-learning/confusion-matrix-machine-learning/

[15] G. Tanzer, M. Shengelia, K. Harrenstien, and D. Uthus, "Context-aware sign language translation," Google Research and Rochester Institute of Technology, 2022.

[16] P. C. Alvarez, X. G. Nieto, and L. T. Benet, "Transformer-based sign language translation for How2Sign dataset," 2022.

[17] M. S. Astriani and M. Alvianto, "Web-based app for remote health consultation and machine learning model for sign classification for COVID-19," Department of Computer Science, Bina Nusantara University, Indonesia, 2021.