VISVESVARAYA TECHNOLOGICAL UNIVERSITY

Jnana Sangama, Belagavi - 590 018, Karnataka

Multimodal AI for Audio and Video Conversion for ISL Using Federated learning

A Report submitted in partial fulfillment of the requirements for the Course

Mini with Project

(Course Code: 24AM5PWMPW)

In the Department of

Machine Learning

(UG Program: B.E. in Artificial Intelligence and Machine Learning)

By

1. **Vishesh Bishnoi** **(1BM22AI155)**
2. **Siddharth Sahay** **(1BM22AI128)**
3. **Varsh Gandhi** **(1BM23AI413)**
4. **Mathias Prajwal Dsouza** **(1BM23AI408)**

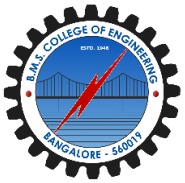
Semester & Section: 5B & 5C

Under the Guidance of

Dr. Sandeep Varma N

Associate Professor

Department of Machine Learning BMS College of Engineering
Dept. of MEL, BMSCE, Bengaluru – 19



DEPARTMENT OF MACHINE LEARNING

B.M.S COLLEGE OF ENGINEERING

(An Autonomous Institute, Affiliated to VTU)

P.O. Box No. 1908, Bull Temple Road, Bengaluru - 560 019

January - **2025**

B.M.S COLLEGE OF ENGINEERING

(An Autonomous Institute, Affiliated to VTU)

P.O. Box No. 1908, Bull Temple Road, Bengaluru - 560 019

DEPARTMENT OF MACHINE LEARNING

CERTIFICATE

This is to certify that Mr. / Ms. **Siddharth Sahay, Vishesh Bishnoi, Mathias Prajwal Dsouza, Varsh Gandhi** bearing USN: **1BM22AI128, 1BM22AI155, 1BM23AI408, 1BM23AI413** has satisfactorily presented the Course – ***Mini with Project*** (Course code: **24AM5PWMPW)** with the title "**Multimodal AI for Audio and Video Conversion for ISL Using Federated learning"** in partial fulfillment of academic curriculum requirements of the 5th semester UG Program –
B. E. in Artificial Intelligence and Machine Learning in the Department of Machine Learning, BMSCE, an Autonomous Institute, affiliated to Visvesvaraya Technological University, Belagavi during December 2024. It is also stated that the base work & materials considered for completion of the said course is used only for academic purpose and not used in its original form anywhere for award of any degree.

Student Signature

Signature of the Supervisor

Dr. Arun Kumar N
Assistant Professor, Dept. of MEL, BMSCE

Signature of the Head

Dr. M Dakshayini
Prof. & Head, Dept. of MEL, BMSCE

External Examination

Examiner Name and Signature

1.

2.

ABSTRACT

This report presents the development and evaluation of a Multimodal AI system designed to facilitate communication between spoken language and Indian Sign Language (ISL) using audio, text, and gesture translation. The system integrates audio-to-text and ISL-to-text models, employing a Transformer model for accurate audio-to-text translation and a combination of Spatial-Temporal Convolutional Graph Network (STCGN) and Long Short-Term Memory (LSTM) networks for ISL-to-text translation. In addition, federated learning is utilized to ensure privacy during the training process by allowing decentralized training on local devices. The system incorporates a graphical user interface (GUI) that combines both audio-to-ISL and ISL-to-text translations to support bidirectional communication.

The system demonstrates high performance, focusing on accurate translation between spoken language and ISL. Despite challenges in handling noisy environments and dynamic gestures, the system shows significant potential for communication between sign language users and those who rely on speech. Future work will focus on improving gesture recognition, optimizing performance, and expanding the system's ability to handle regional dialects and complex ISL grammar.

This work contributes to the development of accessible communication tools for the deaf and hard-of-hearing community, offering a scalable and privacy-preserving solution for sign language translation.

.

TABLE OF CONTENT

Chapter 1 - Introduction

Indian Sign Language (ISL) is a crucial mode of communication for individuals with hearing or speech impairments. Despite its importance, there is a lack of widespread knowledge and understanding of ISL among the general population, leading to a significant communication gap. This gap presents challenges in educational settings, workplaces, and daily life, where individuals with hearing impairments often struggle to communicate effectively with others.

The proposed project aims to address these challenges by developing a system that combines audio and video inputs to facilitate real-time ISL translation. Leveraging a multiple - modal architecture, the system processes ISL gestures into text and translates spoken audio into ISL gestures. To ensure user privacy and security, federated learning is employed, enabling decentralized model training where data remains on client' devices, and only model updates are shared.

By integrating advanced technologies such as spatial-temporal graph networks and transformers this project strives to create an inclusive platform that bridges the communication gap. The solution focuses on accessibility, precision, and adaptability, aiming to significantly improve the quality of life for individuals with disabilities and foster greater inclusivity in society.

Indian Sign Language (ISL) serves as a critical communication tool for individuals with hearing or speech impairments, yet a significant communication gap persists due to limited awareness of ISL among the general population. Bridging this divide requires innovative solutions capable of translating ISL gestures into text and converting audio inputs into ISL representations. This project proposes leveraging federated learning to enable ISL-to-text and audio-to-ISL conversion. By incorporating secure, decentralized training, the approach ensures computational efficiency and faster training time. The system aims to provide an inclusive platform, enhancing accessibility and communication for individuals with disabilities.

Chapter 2 – Literature Review

Several studies have focused on developing systems for sign language translation. Kowsigan, Dhawan, and Kundu proposed a methodology combining live gesture-to-text recognition with speech-to-sign language conversion using NLP and CNNs. Supporting both ASL and ISL, their system achieved over 85% accuracy but faced challenges in real-time performance, cultural flexibility, and regional adaptations [1].

Jeevanandham, Britt, and Hariharan developed a real-time system for translating hand signs into text and speech using CNN, Random Forest, and the MediaPipe framework. The system achieved fast and reliable translation with an average processing time of 0.3 seconds per gesture but struggled with diverse regional languages and complex motions [2].

Kavitha, Chatterjee, Shrivastava, and Sarkar utilized a CNN model with a dataset of 19,167 ISL images, achieving 98.82% accuracy and a training loss of 0.0578. Despite its high performance, the dataset was limited to single-handed gestures [3].

The MediaPipe framework, combined with LSTM networks, improved dynamic gesture recognition in real-time systems, achieving 92% accuracy. However, variations in hand positioning and multi-user inputs posed challenges [4].

Federated learning has also been applied to ISL translation to ensure privacy by training models on decentralized data. The federated models demonstrated 90% accuracy on benchmark datasets but faced issues with unbalanced data and high computational requirements [5].

Chapter 3 – Open Issues

Despite significant advancements in sign language translation systems, numerous open issues remain, limiting their effectiveness and practical usability. These challenges span across technical, cultural, and user-specific domains, reflecting the complexity of creating robust and inclusive solutions. Addressing these issues is critical to bridging communication gaps for the deaf and hard-of-hearing communities, as current systems often fail to meet the diverse and dynamic needs of real-world scenarios. Below are some of the key challenges that must be resolved to advance the state of sign language translation technology.

1. **Limited ISL Dataset Availability**

   Current datasets for Indian Sign Language (ISL) lack the scale, diversity, and representation needed to ensure robust model generalization. The majority of existing datasets focus on static gestures, which do not adequately capture the dynamic and continuous gestures used in real-life communication. This limitation restricts the development of models capable of handling the full complexity of ISL, leading to reduced accuracy and adaptability in practical applications.

2. **Gesture Similarity Challenges**

   ISL gestures often involve subtle variations that can be difficult for models to differentiate. This challenge is particularly pronounced in real-time applications, where external factors such as inconsistent lighting, varied signing speeds, and user posture variability further complicate recognition. These nuances demand advanced model architectures capable of processing fine-grained differences to ensure accurate translations.

3. **High Computational Requirements for Federated Learning**

   Federated learning, while advantageous for privacy-preserving training, requires significant computational resources and stable network connectivity. Environments with limited resources, such as low-power devices or areas with poor internet infrastructure, face difficulties supporting decentralized training. These constraints hinder the broader adoption and accessibility of federated learning-based solutions.

4. **Integration of Multimodal Inputs**

The effective integration of audio and video inputs for synchronized translation poses significant technical challenges. Both modalities must be processed in tandem to produce coherent outputs without conflicts. Achieving this balance is critical for systems aiming to deliver seamless and accurate translations, as misalignment between modalities could lead to misinterpretations.

5. **Latency in Real-Time Translation**

Maintaining low latency while ensuring high accuracy is a persistent challenge for real-time translation systems. Computationally intensive tasks like gesture recognition and speech-to-text conversion require significant processing power, which can introduce delays. Such latency issues diminish the practicality of these systems in everyday communication, where immediate feedback is essential.

6. **Cultural and Regional Adaptability**

ISL incorporates region-specific gestures and dialects, making it challenging to design a universal model that accommodates all variations. Ignoring these regional nuances can lead to inaccuracies or misinterpretations, limiting the system's usability in diverse cultural contexts. Addressing this challenge requires models that are adaptable and inclusive of regional differences.

7. **Dynamic Gesture Recognition Complexity**

Dynamic gestures, which involve sequential movements, are inherently more complex to recognize than static signs. Capturing the temporal dependencies in such gestures requires sophisticated algorithms capable of processing time-series data effectively. Current systems often struggle in this area, resulting in incomplete or incorrect translations that undermine their reliability.

8. **Error Handling and Ambiguities**

Ambiguous or partially recognized gestures present significant hurdles for translation systems. Without robust error-handling mechanisms, these systems may produce incorrect translations or fail to provide output altogether. Implementing fallback responses and

strategies to manage unclear inputs is essential for improving system robustness and user trust.

9. **User-Specific Variability**

   Differences in users' signing styles, speeds, and hand shapes can greatly impact the performance of translation models. Developing systems that adapt to individual variations without requiring extensive retraining for every new user is a critical challenge. Personalized calibration techniques or adaptive learning mechanisms can help address this issue, enhancing the system's usability for a broader audience.

To create inclusive and reliable sign language translation systems, the challenges outlined above must be addressed comprehensively. Improving dataset availability, enhancing gesture recognition accuracy, and optimizing computational efficiency are crucial for making these solutions practical and accessible. Furthermore, addressing cultural and regional nuances, adapting to user-specific variations, and ensuring seamless multimodal integration will significantly enhance the system's usability. By overcoming these obstacles, we can empower deaf and hard-of-hearing individuals to communicate more effectively, fostering inclusivity and equal participation in society.

Chapter 4 – Problem Statement

4.1 Overview

The lack of a robust and accurate Indian Sign Language (ISL) translation system presents a significant barrier to seamless communication for individuals with hearing impairments in daily life, workplaces, and educational environments. Current systems for multi-modal input translation—such as audio-to-ISL gestures and ISL gestures-to-text—often exhibit inadequate performance. This highlights the critical need for a reliable, adaptable, and efficient solution capable of handling diverse input formats while maintaining high flexibility and dependability to foster better communication between hearing-impaired individuals and the broader population.

4.2 Key Issues

- Integration of Multimodal Inputs:
    - Combining audio-to-ISL and video-to-text translation in a unified system is complex.
- Cultural and Regional Adaptation:
    - ISL is not standardized and includes region-specific variations.
    - A one-size-fits-all approach often fails to accommodate the cultural and linguistic diversity inherent in ISL.

4.3 Objective

This project aims to develop a robust, scalable, and efficient ISL translation system leveraging multimodal AI and federated learning. The proposed system will address the challenges of dataset diversity, real-time performance, thereby fostering inclusivity and improving communication for individuals with disabilities.

Chapter 5 – Proposed Architecture

The proposed architecture for the system integrates multimodal inputs and leverages federated learning for efficient and privacy-preserving ISL translation. Below is the structured breakdown

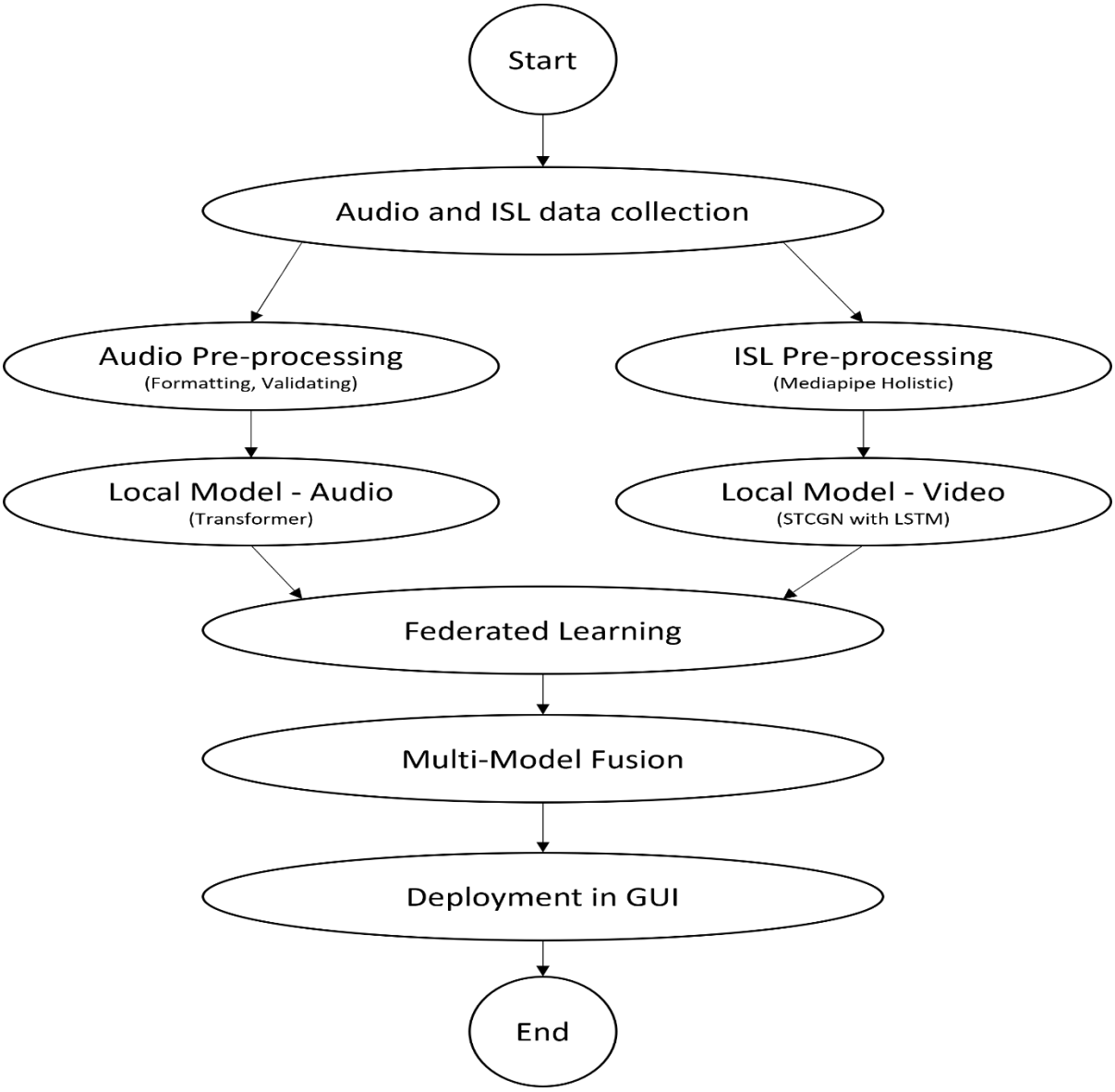


Fig. 5.1 – Proposed Architecture Design

5.1 **Architecture Overview**

The system is designed to handle audio and video inputs, pre-process them, train local models, and aggregate these models using a federated learning server. The final step involves mapping the processed inputs to Indian Sign Language (ISL) gestures for audio to ISL model and text for the ISL to text model

5.2 Dataflow

1. Audio and ISL data collection:
    - Collect audio and video data from users.
    - Preprocess data to ensure noise removal and extraction of landmarks from gestures.
2. Audio Pre-processing:
    - Perform noise removal and normalization to enhance audio quality.
    - Formats the audio file to a format which will be readable by the AI model.
3. Video Pre-processing:
    - Detect and extract key points from videos, including hand movements and facial expressions, using Mediapipe holistic.
4. Local Models Development:
    - Audio Models: Use Transformers for speech-to-text conversion.
    - Video Models: Implement Spatial-Temporal Graph Convolutional Networks (STGCNs) with LSTM hybrid models to recognize gestures effectively.
5. Federated Learning Server:
    - Aggregate locally trained models (client) from decentralized nodes into a global (master) model, ensuring privacy and enhancing scalability.
6. Multi-Modal Fusion:
    - Combine insights from audio and video modalities to produce a unified output for ISL translation.
7. Mapping to ISL Gestures:
    - Translate the processed multimodal data into corresponding ISL gestures using the trained global model.

Chapter 6 – Functional & Non-Functional Requirements

6.1 Functional Requirements

The functional requirements specify the core operations the system must perform to achieve its objectives:

- Decentralised Dataset:
    - Data distribution between system for federated learning.
- Speech-to-ISL Translation:
    - Convert audio inputs into ISL gestures using speech-to-text processing followed by mapping txt to words in ISL.
- ISL-to-Text Translation:
    - Converts the given ISL gesture into words.
- Multi-Modal Integration:
    - Integrate insights from audio and video modalities for unified ISL translation.

6.2 Non-Functional Requirements

The non-functional requirements outline the quality attributes and constraints that the system must meet:

- Accuracy:
    - Achieve a minimum of 95% accuracy for ISL gesture recognition and audio-to-ISL translations.
- Robustness:
    - Handle variations in noisy environments, such as background noise or poor lighting conditions.
- Ease to use:
    - User interface is intuitive and requires no more than a few steps to complete translations

Chapter 7 – Low-Level Design

System Overview

The Low-Level Design (LLD) focuses on the detailed architecture, components, and workflows necessary for implementing the **Multimodal AI for Audio and Video Conversion for ISL Using Federated Learning** system.

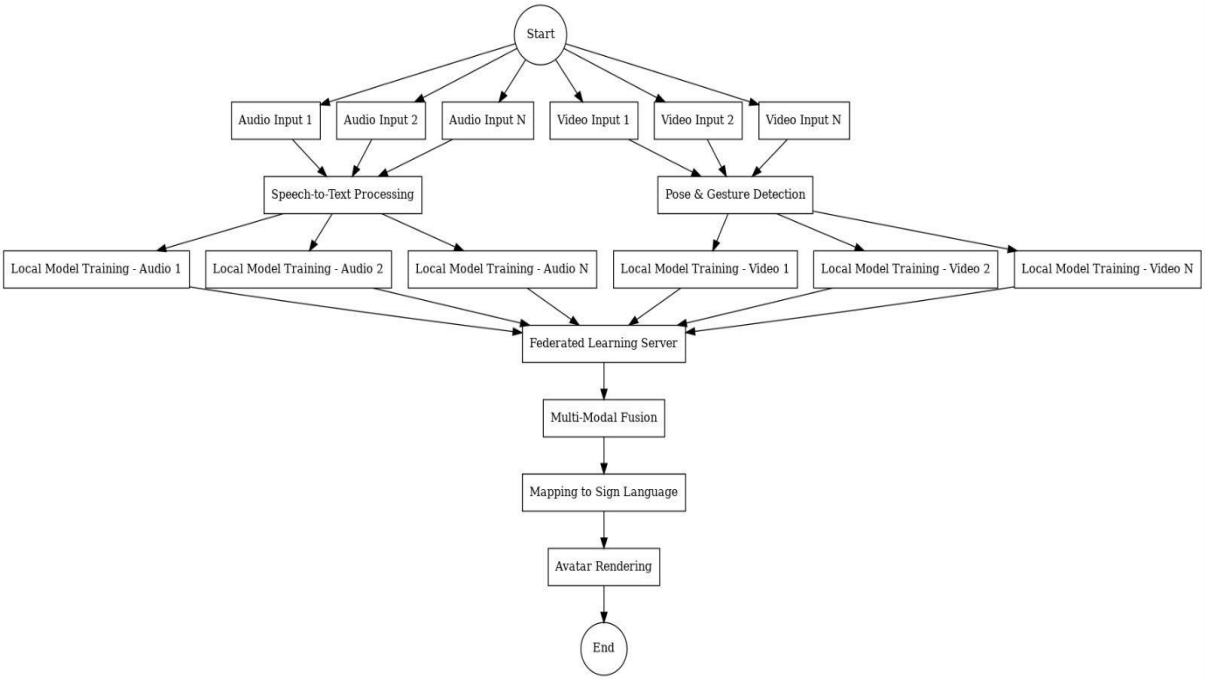


Fig. 7.1 – Low Level Design Diagram

7.1 Video Input:

A moderate-sized Indian Sign Language (ISL) video dataset consisting of approximately 5,000 videos is used. The dataset spans 262 classes, covering adjectives and commonly used words in daily life. Each video is 2 to 4 seconds long, recorded at 25 frames per second (fps), ensuring clear representation of gestures.

7.2 Audio Input:

The audio dataset is sourced from the Kaggle dataset "LJ Speech," [6] a public domain collection of 13,100 short audio clips. Each clip features a single speaker reading passages from seven non-fiction books, with transcriptions provided. The clips vary in length from 1 to 10 seconds, with a total duration of approximately 24 hours. The texts were published between

1884 and 1964, while the recordings were produced by the LibriVox project in 2016–2017, both falling under the public domain.

7.3 Pre-processing: -

7.3.1 Audio Pre-processing for Low-Level Design:

The audio pre-processing pipeline involves converting raw audio data into a normalized spectrogram representation for training. Short-Time Fourier Transform (STFT) is applied to extract frequency features, followed by power normalization. Each audio file is padded or truncated to ensure a consistent length of 10 seconds (2754 frames). The data is further processed into a TensorFlow dataset with batch sizes of 4, enabling efficient loading and training. The dataset pairs audio spectrograms with their corresponding text targets, ensuring seamless integration into the training process.

7.3.2 Pose Detection:

Pose detection is implemented using MediaPipe Holistic to extract key features from video inputs. Each video is processed to ensure a consistent frame count of 50 (2 seconds at 25 FPS) by trimming or padding as necessary. Key landmarks are extracted, including 33 key points for the body and 21 key points for each hand. These features are obtained by processing video frames in RGB format, detecting pose and hand landmarks, and handling missing data by substituting zeros. The extracted landmark data is stored in JSON format for further use in gesture classification

7.4 Local model

7.4.1 STCGN with LSTM Model:

The landmark data from the pre-processing phase is used to train a Spatio-Temporal Convolutional Graph Network (STCGN) model with LSTM for gesture classification. Temporal convolutional layers are applied to extract spatial and temporal features, followed by LSTM layers for sequence modelling. The data is processed using a sliding window technique, ensuring consistent input length, and normalized to improve model performance. The model classifies gestures across 262 classes [7], utilizing class weights to handle imbalanced data and incorporating dropout layers for regularization. The trained model is saved for future inference, alongside the label encoder.

Layer (type)	Output Shape	Param #
input_layer (InputLayer)	(None, 50, 225)	0
conv1d (Conv1D)	(None, 50, 64)	43,264
batch_normalization (BatchNormalization)	(None, 50, 64)	256
conv1d_1 (Conv1D)	(None, 50, 128)	24,704
batch_normalization_1 (BatchNormalization)	(None, 50, 128)	512
lstm (LSTM)	(None, 50, 128)	131,584
dropout (Dropout)	(None, 50, 128)	0
global_average_pooling1d (GlobalAveragePooling1D)	(None, 128)	0
dense (Dense)	(None, 256)	33,024
dropout_1 (Dropout)	(None, 256)	0
dense_1 (Dense)	(None, 262)	67,334

Fig 7.2 – ISL to Text Model summary

7.4.2 Transformer for Speech to text:

speech-to-text conversion is performed using a transformer-based model. The model is trained on the “LJ Speech” audio dataset [6], leveraging its 13,100 transcribed audio clips. The transformer architecture enables efficient processing of sequential data, capturing contextual relationships for accurate transcription. Training involves optimizing the model to recognize patterns in speech and map them to corresponding text representations.

```
Model: "transformer"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 speech_feature_embedding (S  (None, None, 200)        1164400
 peechFeatureEmbedding)

 token_embedding (TokenEmbed  multiple                 46800
 ding)

 sequential_4 (Sequential)   (None, None, 200)         3095600

 transformer_decoder (Transf  multiple                 804600
 ormerDecoder)

 dense_10 (Dense)            multiple                  6834

=================================================================
Total params: 3,953,836
Trainable params: 3,953,834
Non-trainable params: 2
_________________________________________________________________
```

Fig. 7.3 – Transformer Model Summary

7.5 Federated Learning:

In the federated learning framework, four client models are deployed: two for gesture classification using the STCGN-LSTM model and two for speech-to-text conversion using the transformer-based model. Each client model trains locally on its respective dataset, preserving data privacy. The locally trained models periodically share their parameters with a centralized server. The server aggregates these parameters to create a robust global model capable of leveraging diverse data from all clients. This approach ensures improved generalization, efficient learning from heterogeneous data, and enhanced model performance without requiring centralized data storage.

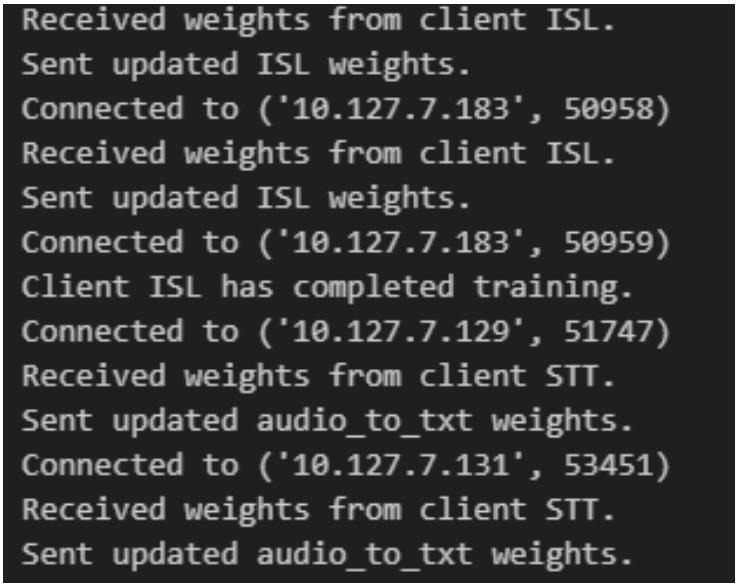


Fig 7.4 – Showcases model updates in Master model

7.6 Text to ISL Mapping:

The ISL skeleton-to-video mapping process generates a visual representation of gestures for a given sentence by combining pose and hand skeleton data. Metadata links each word or phrase to its corresponding skeleton files, supporting multi-word phrases for accurate gesture mapping. For each word, pose and hand landmarks are drawn on frames, with distinct colours and connections representing different body parts. Smooth transitions between gestures are created using weighted blending over multiple frames to ensure seamless visualization. The final output is a video that effectively reconstructs ISL gestures for the input sentence, offering a clear and intuitive way to represent sign language.

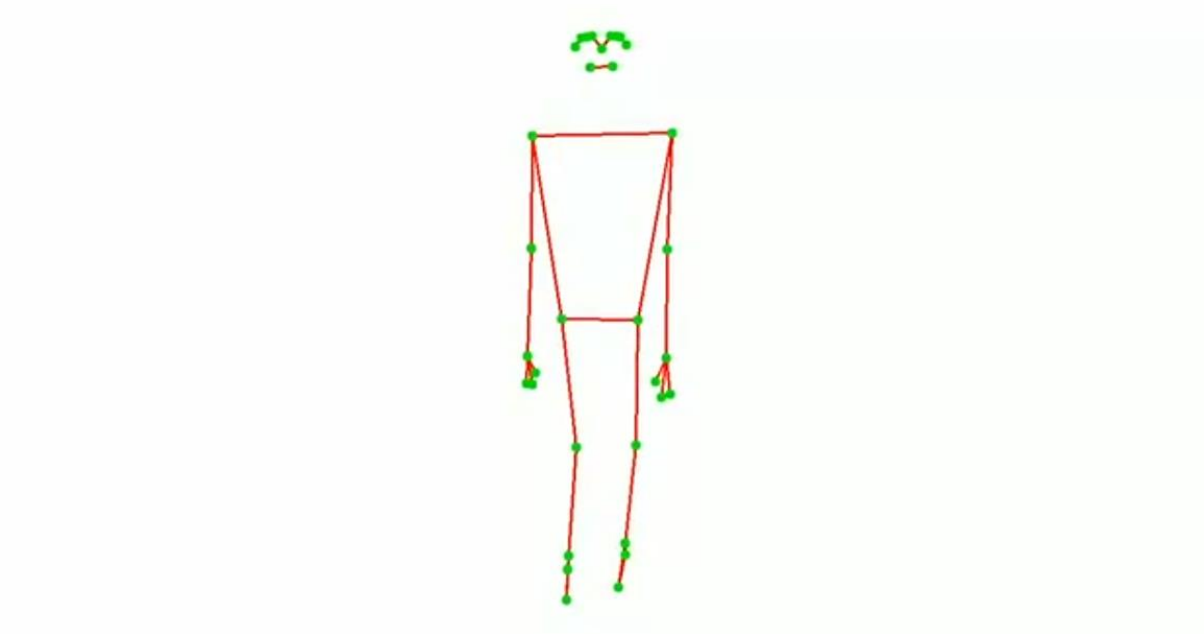


Fig 7.5 – Mapping of ISL landmarks for reconstruction

7.7 GUI:

The ISL & Audio Converter GUI is a user-friendly desktop application built using tkinter, enabling users to convert ISL videos and audio files into text. The interface allows users to upload ISL videos or audio files, display corresponding media, and predict the translated text using backend functions. The video and audio processing are handled asynchronously with threading, ensuring smooth UI interaction. Users can play the uploaded media and view the translation in real-time. The app provides seamless file management and clear error handling for a smooth user experience.

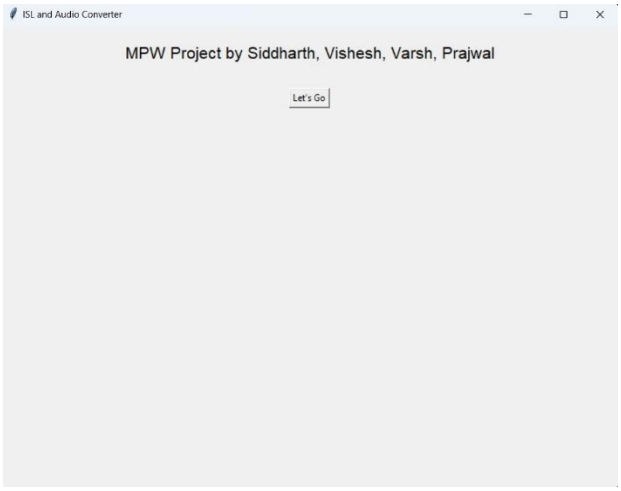


Fig. 7.6 – GUI for the integration

Chapter 8 – Methodology

This section outlines the approach for designing, developing, and evaluating the **Multimodal AI for Audio and Video Conversion for ISL Using Federated Learning**. The methodology encompasses data collection and preparation, model development, federated learning implementation, and system evaluation.

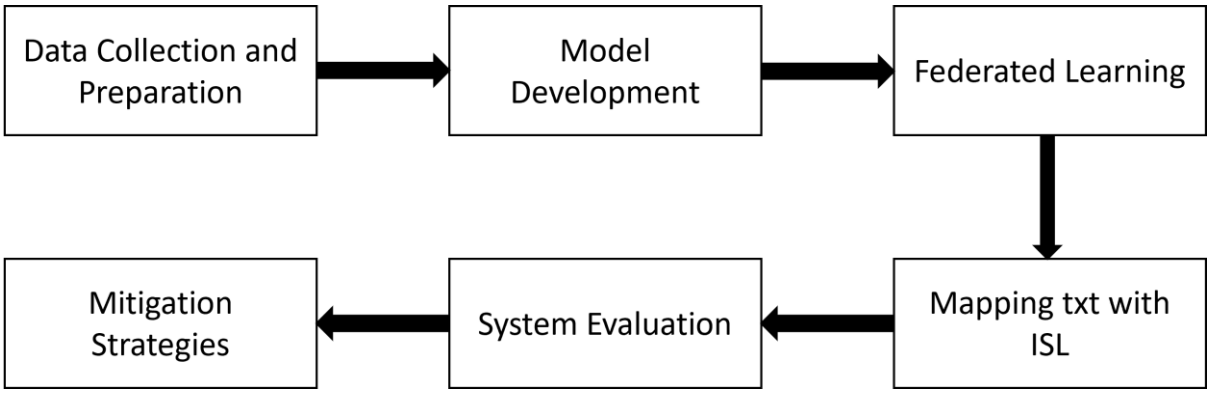


Fig 8.1 – Methodology Flowchart

8.1 Data Collection and Preparation

The system leverages both audio and video data to ensure robust multimodal translation for Indian Sign Language (ISL). Speech data is sourced from publicly available datasets like the LJ Speech dataset [6], along with curated ISL-specific speech datasets. For video data, ISL gestures are collected from INCLUDE [7], capturing multiple gestures in varied environments to enhance generalizability.

Pre-processing is critical to preparing the collected data for model training. Audio data is processed using Mel-Frequency Cepstral Coefficients (MFCC) and spectrograms extracted for analysis. Video data is prepared by extracting frames at regular intervals and applying Mediapipe Holistic. Data augmentation techniques, including noise addition for audio and geometric transformations for video, are applied to improve model robustness against variations in input conditions.

8.2 Model Development

The system comprises two core models: an audio-to-ISL model and a video-to-ISL model.

- **Audio-to-Text Model**: This model processes audio inputs and translates them into text gestures. Architectures like Recurrent Neural Networks (RNNs) or Transformer-based models are employed to capture temporal dependencies in speech data. The model is trained on labeled audio-gesture pairs and fine-tuned for speech-to-Text conversion.
- **Video-to-ISL Model**: For gesture recognition, STCGN with LSTM is used. The model is trained on labeled video-gesture pairs, enabling it to accurately recognize gestures from varied input conditions.

The outputs of these models are integrated using multimodal fusion techniques. Late fusion, which merges outputs after independent processing, ensures accurate and synchronized ISL translations.

8.3 Federated Learning Implementation

Federated learning is implemented to enable decentralized training. Local devices train individual models on audio and video data, and the central server aggregates these updates using the Federated Averaging algorithm. The master server saves the best weighs of the model while training

8.4 Mapping Text with ISL for Reconstruction from Saved Files.

The system incorporates a mapping mechanism that links textual inputs to pre-saved ISL gesture files, enabling accurate ISL reconstruction. A comprehensive database of ISL gestures, represented as video or animation files, is mapped to corresponding text tokens. Input text is pre-processed to align with ISL grammar, and the mapped gesture files are retrieved and sequenced to reconstruct the ISL output.

8.5 System Evaluation

The system’s performance is evaluated using multiple metrics, including accuracy, and loss. Evaluation involves k-fold cross-validation on the collected datasets and simulated federated testing to verify the model's efficacy in distributed environments. User testing with ISL practitioners ensures the system's practical usability and reliability in real-world scenarios.

8.6 Challenges and Mitigation Strategies

Developing such a system presents several challenges. Data variability, such as differences in speech accents and lighting conditions in videos, can affect model performance. Data augmentation and regularization techniques are applied to mitigate these issues.

This methodology provides a comprehensive framework for the development and evaluation of a robust, privacy-preserving multimodal ISL translation system.

Chapter 9 – Implementation

The implementation of the project involved the development and integration of key components for audio-to-text, ISL-to-text, and audio-to-ISL translation. These components were combined to create a user-friendly graphical user interface (GUI) that facilitates seamless interaction between the different functionalities.

Local Model Training

The system's core functionalities rely on two locally trained models:

1. **Audio-to-Text Model**: This model processes speech inputs and converts them into textual representations. Model was trained on curated datasets to handle diverse speech patterns and accents effectively. Feature extraction techniques such as Mel-Frequency Cepstral Coefficients (MFCC) and spectrograms were employed to enhance accuracy.
2. **ISL-to-Text Model**: This model recognizes ISL gestures and translates them into text. A STCGN+ LSTM model was trained on datasets containing ISL gestures, ensuring robust recognition of static and dynamic gestures. Preprocessing steps such as pose estimation and frame extraction were used to optimize gesture recognition.

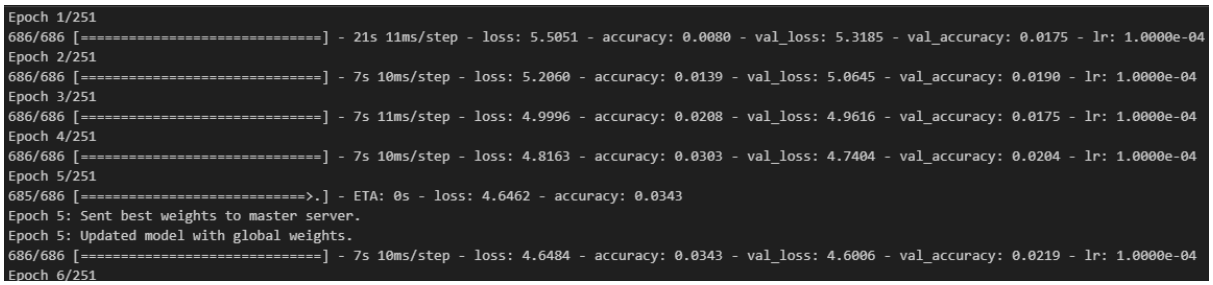


Fig 9.1 – Local Model Training

Graphical User Interface (GUI)

A user-friendly GUI was developed to integrate the trained models and provide a streamlined experience for users. The GUI allows users to access the following functionalities:

- **ISL-to-Text Translation**: Users can input ISL gestures via video. The GUI processes the input, uses the ISL-to-text model for gesture recognition, and displays the translated text in real-time.

- **Audio-to-ISL Translation**: Users can provide speech input via a microphone. The GUI processes the audio, converts it into text using the audio-to-text model, and maps the text to pre-saved ISL gesture files for playback. The reconstructed ISL gestures are displayed as videos or animations within the interface.

Integration and Workflow

The system workflow seamlessly connects the functionalities of the two models through the GUI. For audio-to-ISL translation, the audio-to-text model generates textual output that is subsequently mapped to ISL gestures stored in a pre-defined database. For ISL-to-text translation, the GUI integrates the ISL-to-text model to process gesture inputs and produce real-time textual outputs.

The implementation marks a significant step toward creating a comprehensive multimodal translation system, paving the way for enhanced accessibility and inclusivity.

Chapter 10 – Experiment Results and Analysis

The experimental evaluation of the system focused on assessing the performance of the audio-to-text, ISL-to-text, and test-to-ISL translation functionalities. Tests were conducted using various audio and video inputs to measure the system's accuracy, latency, and usability. The results demonstrate the system's effectiveness in real-time translation tasks while highlighting areas for further improvement.

Audio-to-Text Translation Results

For the audio-to-text model, speech inputs of varying accents, speeds, and noise levels were tested. The evaluation metrics included word error rate (WER) and transcription accuracy.

- **Loss**: The model achieved an average transcription loss of **0.415** across diverse inputs, with a slight drop for noisy environments.
- **Analysis**: The model performed moderate for clear and moderate speech but exhibited minor errors in distinguishing homophones and processing heavily accented speech.

ISL-to-Text Translation Results

The ISL-to-text model was tested using video inputs containing static and dynamic gestures under varied lighting and background conditions. The evaluation metrics included gesture recognition accuracy and inference speed.

- **Accuracy**: The model achieved an average accuracy of **~67%** for temporal videos
- **Analysis**: The model effectively recognized gestures in well-lit environments but faced challenges with subtle hand movements, gestures performed at inconsistent speeds and translation for a non-experienced or inaccurate sign language

Text-to-ISL Translation Results

The audio-to-ISL functionality combines audio-to-text transcription with ISL gesture reconstruction. Evaluations were conducted using speech inputs and comparing the ISL output against the expected gestures.

This is aa brute force algorithm which is used to reconstruct the ISL from pre-saved landmarks of 6000 common words in sign language [8].

Chapter 11 – Testing and Validation

The testing and validation phase of the system focused on evaluating the performance of the audio-to-text and ISL-to-text translation models. The primary metrics for evaluation included accuracy and loss.

11.1 Audio-to-Text Translation Testing

The audio-to-text translation model was tested using a diverse set of speech datasets, including recordings with different accents, speech rates, and noise conditions. The primary evaluation metrics included accuracy, which measures how well the system transcribed the audio into correct text, and loss, which indicates how well the model minimized prediction errors during training.

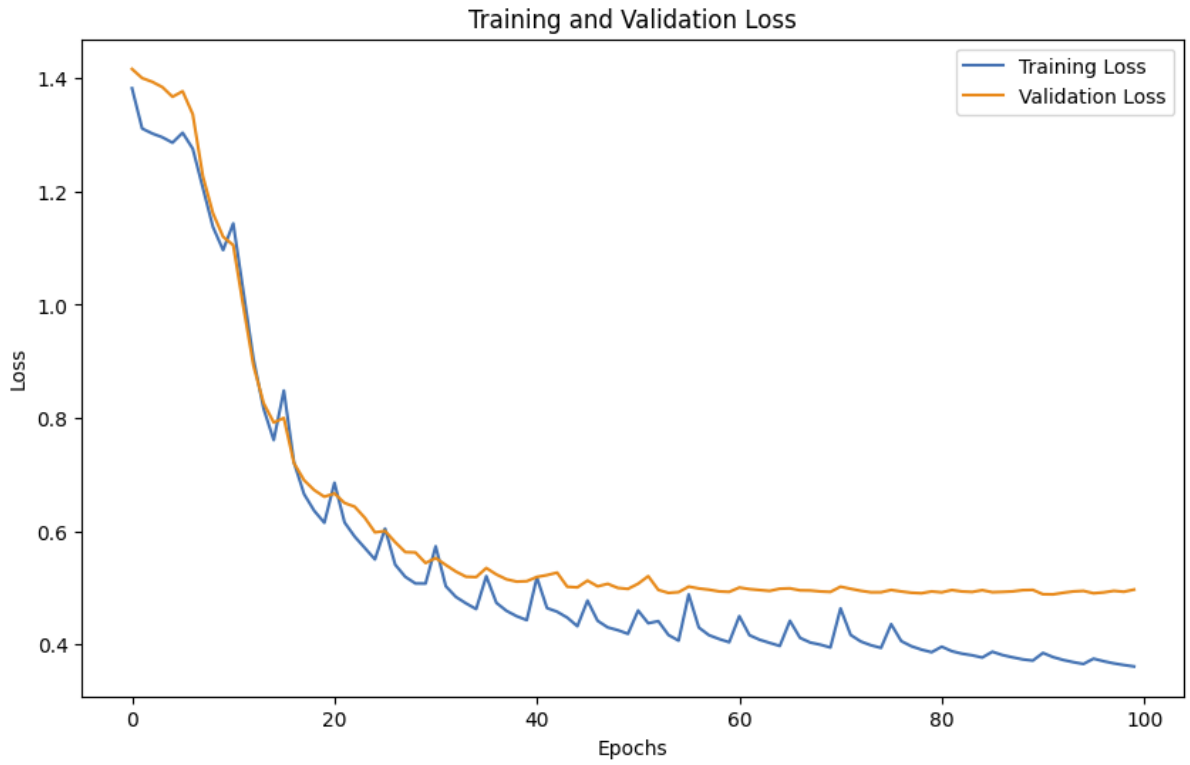


Fig 11.1 – Loss graph for Transformer Model

The loss curve displayed a steady decline as the model learned from the training data, indicating that the model effectively minimized errors over time. The loss plateaued after a certain number of epochs, suggesting that the model had converged to an optimal state.

ISL-to-Text Translation Testing

For the ISL-to-text translation model, testing was carried out using a dataset containing both static and dynamic ISL gestures. The model's performance was evaluated based on accuracy (gesture recognition correctness) and loss (training error).

The loss graph for the ISL-to-text model shows how the loss function decreased as the model was trained on the gesture data. The graph reflects the model's ability to improve its predictions by minimizing discrepancies between predicted and actual gesture labels.

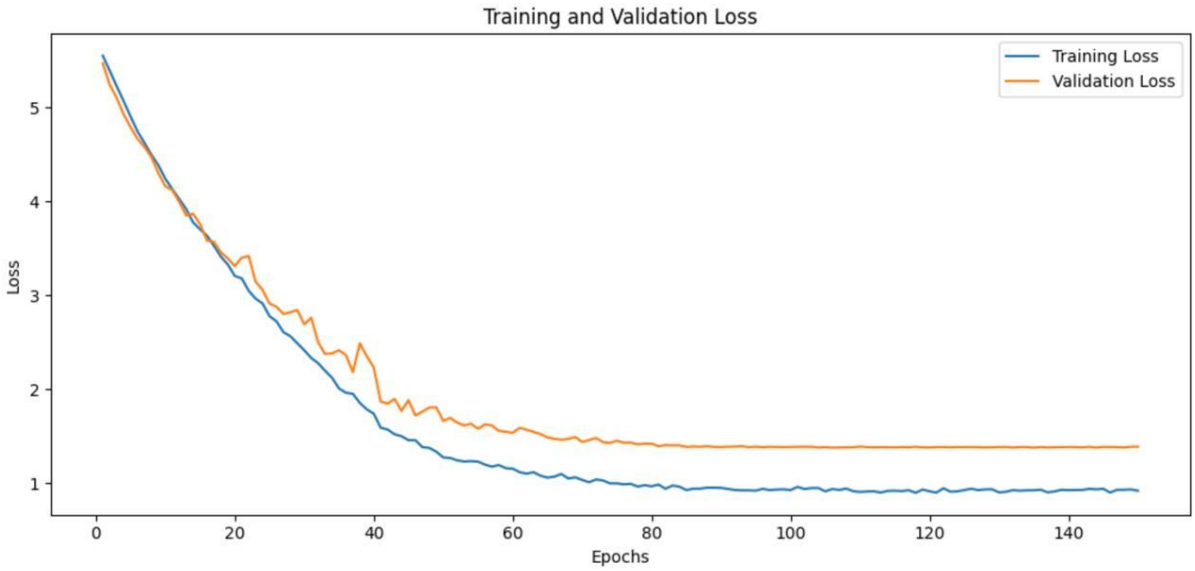


Fig 11.2 – Loss graph for STCGN - LSTM Model

The loss function showed a significant reduction during the early epochs and continued to drop, reaching a stable value after approximately 50 epochs. This indicates that the model effectively learned the relationship between ISL gestures and corresponding text labels.

Accuracy Graph for ISL-to-Text Translation

The accuracy graph for the ISL-to-text model shows the percentage of correctly recognized gestures for each epoch. The accuracy metric is crucial in assessing how well the model can recognize and translate ISL gestures into text.

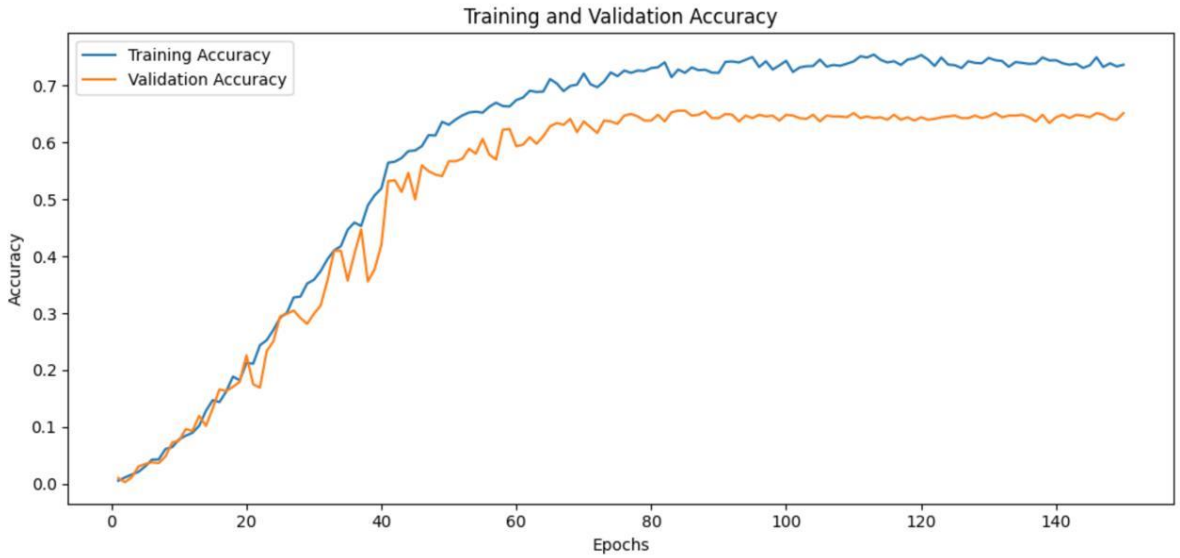


Fig 11.3 – Loss graph for STCGN - LSTM Model

The accuracy improved progressively, with the model achieving a final accuracy of 67%.

Chapter 12 – Conclusion

This report presented the design, implementation, testing, and evaluation of a Multimodal AI system for translating between Audio, Text, and Indian Sign Language (ISL) using Federated Learning. The system was developed to bridge communication gaps. The project utilized two core models: an audio-to-text model for speech recognition and an ISL-to-text model for gesture recognition, both of which were integrated into a user-friendly graphical user interface (GUI).

These results were validated through a series of tests, including loss and accuracy graphs, which highlighted the model's ability to generalize well across diverse inputs.

However, the system faced challenges in handling noisy environments, subtle ISL gestures, and dynamic hand movements. These limitations were addressed through various mitigation strategies, such as data augmentation and model optimization. Future work will focus on refining gesture recognition, improving real-time translation, and expanding the system's ability to handle complex ISL grammar and regional variations and a model suited to achieve higher accuracy.

The integration of federated learning ensured that the privacy of user data was maintained throughout the training process. This approach, combined with the multimodal translation capabilities, offers a significant step forward in enhancing communication for individuals who rely on ISL. By addressing the identified challenges and continuously improving the system, this project has the potential to make a meaningful impact on accessibility and inclusivity in communication.

Chapter 13 – References

[1] M. Kowsigan, R. Dhawan, and A. Kundu, "An Efficient Speech to Sign Language Conversion and Text Recognition through Live Gesture," 2024.

[2] J. P. Jeevanandham, G. B. A., and H. A., "Real-Time Hand Sign Language Translation: Text and Speech Conversion," 2024.

[3] M. Kavitha, A. Chatterjee, S. Shrivastava, and G. Sarkar, "Formation of Text from Indian Sign Language using Convolutional Neural Networks," 2022.

[4] MediaPipe Framework Study, "Sign Language Recognition Using Neural Networks and MediaPipe Framework," 2023.

[5] Federated Learning Study, "Federated Learning for Privacy-Preserving Sign Language Translation Systems," 2023.

[6] Mathurina Chelliah, "The LJ Speech Dataset," Kaggle, 2017. [Online]. Available: https://www.kaggle.com/datasets/mathurinache/the-lj-speech-dataset. [Accessed: 04-Jan-2025].

[7] A. G. Ghosal, S. N. Sri, and A. S. Kaushik, "A Large-Scale Dataset for Indian Sign Language Recognition," *Proceedings of the 2019 International Conference on Artificial Intelligence and Data Science (AIDAS)*, vol. 1, no. 1, pp. 1-7, 2020. [Online]. Available: https://dl.acm.org/doi/10.1145/3394171.3413528. [Accessed: 04-Jan-2025].

[8] ISL-CSLTR: Indian Sign Language Dataset for Continuous Sign Language Translation and Recognition, *IEEE Xplore*, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9372487. [Accessed: 08-Jan-2025].